TOWARDS MULTIMODAL ANALYSIS OF HUMAN INTERACTION PATTERNS IN GROUP CONTEXT

KANGZHONG WANG

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University
Department of Computing

Towards Multimodal Analysis of Human Interaction Patterns in Group Context

Kangzhong Wang

A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
November 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgment has been made in the text.

Signature: _____

Name of Student: Kangzhong Wang

Abstract

Understanding how humans communicate, respond, and learn in social and educational settings is central to developing human-aware artificial intelligence. This thesis investigates how multimodal behavioural signals—visual, acoustic, and linguistic—can be modelled to interpret key aspects of human interaction, including responsiveness, agreement, and learning outcomes. Through three consecutive studies, it establishes a computational framework for analysing interactional behaviour across conversational and learning contexts.

The first study focuses on backchannel detection in group conversations, where listeners' subtle nonverbal cues such as head nods and facial movements indicate attentiveness and understanding. A temporal visual attention framework is proposed to capture both micro- and macro-level temporal dependencies across behavioural channels. Experiments on the MPIIGroupInteraction and CCDb datasets demonstrate that the framework effectively recognises diverse backchannel behaviours and outperforms contemporary baselines, highlighting the importance of temporal reasoning and motion-based feature representation.

Building on this foundation, the second study extends the analysis from behavioural events to attitudinal meaning by estimating agreement intensity from multimodal backchannel segments. Audio and visual modalities are integrated through adaptive temporal fusion to model how listeners express varying degrees of agreement. Results show that combining visual motion cues with prosodic and spectral features could

help improve predictive accuracy. Incorporating speaker diarisation further enhances role-specific modelling, revealing that agreement is often conveyed implicitly through coordinated facial and vocal cues.

The final study applies multimodal modelling to an educational context by predicting students' self-reported learning gains in online service-learning programmes. Using audio, video, and transcript features from synchronous sessions, the analysis identifies behavioural indicators associated with different dimensions of learning outcomes—intellectual, social, civic, and intrapersonal. Screen-content features that capture the semantic richness and structure of instructional materials are found to best predict intellectual learning, while visual participation cues—such as camera usage and body posture—correlate more strongly with social and civic learning. These findings indicate that both communicative behaviour and visible participation contribute to effective collaborative learning.

Overall, the three studies establish a coherent research trajectory from low-level behavioural detection to high-level outcome prediction. Methodologically, the thesis contributes interpretable frameworks for temporal attention and multimodal fusion that generalise across conversational and learning domains. Conceptually, it advances the computational understanding of how responsiveness, agreement, and learning outcomes can be inferred from multimodal behaviour, supporting the development of human-aware AI systems capable of perceiving and facilitating meaningful interaction.

Publications Arising from the Thesis

1. **K. Wang**, X. Zhai, M. Cheung, E. Y. Fu, P. Q. Chen, G. Ngai, and H. V. Leong, "TMAN: A Temporal Multimodal Attention Network for Backchannel Detection", in *Neurocomputing*, p.131605, 2025.

2. **K. Wang**, E. Y. Fu, G. Ngai, and H. V. Leong, "Multimodal Learning Analytics for Predicting Learning Gains in Online Service-Learning Programs", accepted by *2025 12th International Conference on Behavioural and Social Computing (BESC)*

3. **K. Wang**, E. Y. Fu, G. Ngai, and H. V. Leong, "Identifying key learning factors in service-learning programs using machine learning". In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2022, pp. 1312–1317.

Other Publications during PhD Study

1. Y. Huang **K. Wang**, E. Y. Fu., G. Ngai, and Peter H.F. Ng, "CMIS-Net: A Cascaded Multi-Scale Individual Standardization Network for Backchannel Agreement Estimation", manuscript submitted to *Pattern Recognition*.

2. **K. Wang**, Z. Shen, Y. Zhang., M. Cheung, X. Luo, G. Ngai, and E. Y. Fu., "One Size Fits All? A Modular Adaptive Sanitization Kit (MASK) for Customizable Privacy-Preserving Phone Scam Detection", accepted by *Proceedings of the 33rd ACM International Conference on Multimedia (MM'25)*.

3. X. Zhai, Y. Wang, L. Liang, **K. Wang**, F. Pei, and E. Y. Fu, 2025. "Personalized e-learning resource recommendation using multimodal-enhanced collaborative filtering", in *Knowledge-Based Systems*, p.113605.

4. Z. Shen, **K. Wang**, Y. Zhang, G. Ngai, and E. Y. Fu, "Combating Phone Scams with LLM-based Detection: Where Do We Stand? (Student Abstract)", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 28, 2025, pp. 29 487–29 489.

5. T. Zhang, **K. Wang**, K. Jing, Li, G. Li, Q. Li, C. Zhang, and H. Yan, "A ring2vec description method enables accurate predictions of molecular properties in organic solar cells", in *npj Computational Materials*, vol. 10, no. 1, p.

268, 2024.

6. C. Yang, **K. Wang**, P. Q. Chen, M. M. Cheung, Y. Zhang, E. Y. Fu, and G. Ngai, "Multi-mediate 2023: Engagement level detection using audio and video features", in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 9601–9605.

Acknowledgments

I would like to express my sincere appreciation to all the people who have supported and helped me throughout my doctoral study. The following acknowledgments are by no means exhaustive, and I apologize to anyone whose name may be unintentionally omitted.

I am deeply grateful to my supervisors, Prof. Grace Ngai and Prof. Hong Va Leong, for their continuous guidance, encouragement, and advice during the course of my Ph.D. study. Their constructive comments, patience, and trust have been essential to the completion of this thesis and my development as a researcher.

I would also like to thank Dr. Eugene Yujun Fu for his careful supervision, thoughtful discussions, and valuable suggestions throughout my research. His support and high standards of academic work have greatly influenced the way I think and conduct research. I am also thankful to Dr. Youqian Zhang for his helpful discussions and assistance during my study.

I have greatly appreciated the opportunity to work with my research group members—Dr. Yuanyuan Wang, Dr. NG Hiu Fung Peter, Michael Cheung, Peter Qi Chen, Chunxi Yang, Xinwei Zhai, Yuxuan Huang, and Zitong Shen. Their collaboration, dedication, and support have made my Ph.D. experience both productive and enjoyable.

I am thankful to my friends Ting Zhang, Yiran Li, Pangjing Wu, and Jia Ye, whose

companionship and encouragement have supported me through many stages of this journey.

I wish to thank my family—my parents and my sister—for their love, patience, and unconditional support. Their confidence in me has been the greatest motivation to complete this work.

Lastly, I would like to thank myself for continuing this journey even when it was difficult. This thesis marks not only an academic milestone but also a personal one.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human communication is inherently multimodal, encompassing speech, facial expression, gesture, posture, and gaze. Through these rich behavioural channels, individuals express thoughts, intentions, and emotions, forming the foundation for collaboration, learning, and social coordination. Within this broad landscape, *group interaction* occupies a particularly significant role, underpinning teamwork in education, workplaces, and community activities. Understanding how individuals behave, respond, and align with others in such group settings has long been a central question across psychology, linguistics, and computational social science. With the growing availability of large-scale multimodal recordings, computational methods now allow us to examine these fine-grained social signals quantitatively, revealing behavioural dynamics that are difficult to capture through human observation alone.

Recent advances in computer vision, audio processing, and multimodal learning have enabled automatic modelling of human interaction at unprecedented temporal and behavioural resolutions. These developments open new opportunities to study how subtle nonverbal signals—such as listener feedback, facial movements, or vocal prosody—reflect internal cognitive or affective states, including attentiveness, agreement, and engagement. However, analysing multimodal behaviour in real-world group contexts remains

a challenging task. Signals are often subtle, asynchronous, and shaped by conversational roles, social relationships, and contextual factors. Moreover, connecting these observable cues to meaningful outcomes, such as social alignment or learning gains, requires models that can capture temporal dependencies and interpersonal variability.

This thesis addresses these challenges by developing and evaluating computational methods for modelling multimodal interaction patterns in group conversation and collaborative learning contexts. Specifically, it focuses on three research directions that progressively expand the analytical scope—from detecting momentary listener responses, to estimating attitudinal meaning, and finally, to predicting learning outcomes:

- **Video-based Backchannel Detection in Group Conversation:** This study develops visual models to automatically detect listener backchannel behaviours—such as head nods, smiles, and gaze shifts—that signal attentiveness and understanding. By analysing visual cues at multiple temporal scales, it provides a foundation for understanding micro-level behavioural coordination during group communication.

- **Multimodal Agreement Estimation from Backchannel Behaviours:** Building upon the detection of backchannel occurrences, this study estimates the *intensity of agreement* expressed through these listener responses. By integrating audio and visual modalities with temporal fusion and role-aware preprocessing, it models how individuals convey concurrence or divergence within natural conversations.

- **Predicting Learning Gains in Online Service-Learning Programs:** Extending the analysis from attitudinal states to educational outcomes, this study investigates how multimodal conversational behaviour relates to students' intellectual, social, civic, and intrapersonal development. By fusing linguistic, visual, and interactional features from recorded group discussions, it demon-

strates the potential of multimodal learning analytics for evaluating authentic learning processes.

These three research threads form a coherent progression: from recognising low-level communicative signals, to inferring high-level social meaning, and finally to predicting broader developmental outcomes. Together, they provide methodological and empirical insights into how multimodal behaviour reflects cognition, affect, and engagement in real-world group interaction.

1.1 Background and Motivation

1.1.1 Backchannel Behaviours in Group Interaction

Backchannels are brief listener responses—such as head nods, smiles, eyebrow raises, or short vocal acknowledgements—that indicate attentiveness or understanding without interrupting the speaker's turn [102, 77]. In group communication, backchannels are fundamental for maintaining conversational flow, coordinating turn-taking, and reinforcing social cohesion. Their frequency, form, and timing vary across individuals, shaped by cultural norms, personal habits, and situational context. Automatically detecting backchannels presents multiple challenges. These behaviours are often short, low in visual salience, and easily occluded in multi-person settings. Moreover, the same gesture may convey different meanings depending on conversational context and participant role.

Recent studies have explored multimodal and temporal attention models to capture these subtle behaviours, combining spatial representations of facial and body motion with temporal cues. However, most existing approaches are limited to dyadic settings or rely on synthetic annotations, leaving group interaction underexplored. The first part of this thesis addresses these gaps by developing a visual-based approach for

robust backchannel detection in multi-party conversations. The proposed method integrates motion-based feature transformation and temporal attention to identify listener feedback in realistic conditions, providing both a methodological foundation and empirical insight for subsequent modelling.

1.1.2 Agreement Expressed by Backchannel Behaviours

While identifying backchannel occurrences reveals *when* listeners respond, it does not explain *what* these behaviours express. Among their possible functions, the expression of agreement is particularly significant for understanding social alignment and consensus building. Agreement can be conveyed both verbally and nonverbally, often through coordinated multimodal patterns such as head nods, smiles, and subtle changes in vocal prosody. Estimating the intensity of agreement from observed behaviour requires models that can integrate information across modalities and account for speaker–listener dynamics.

Existing studies on agreement recognition have largely focused on verbal transcripts or isolated nonverbal features, often neglecting their temporal interdependence. Moreover, the degree of agreement is rarely treated as a continuous spectrum. In this thesis, agreement estimation is formulated as a regression problem that models the continuous variation of attitudinal alignment. Audio features—including energy, frequency, and spectral descriptors—are combined with visual features such as facial action units and head movements to infer agreement intensity. The approach captures how multimodal feedback evolves over time and differs across roles within the group. This study extends the scope of backchannel analysis from occurrence-level detection to attitudinal interpretation, contributing to a deeper understanding of how nonverbal feedback conveys social meaning.

1.1.3 Predicting Learning Gains from Group Conversation

In educational contexts, group discussion serves as an important medium for collaborative knowledge construction and experiential learning. Service-learning, in particular, integrates academic study with community engagement, offering a rich environment for observing learning as a social and reflective process. Traditional assessment methods often overlook these dynamic and interpersonal dimensions of learning. Multimodal learning analytics provides a data-driven means to bridge this gap by linking behavioural indicators with learning outcomes.

This thesis investigates how features derived from group conversation—including speech participation, facial expressivity, and linguistic complexity—relate to students' self-reported learning gains. The study focuses on four major outcome dimensions: intellectual, social, civic, and intrapersonal growth. Through regression-based modelling, it identifies which behavioural indicators correlate with specific types of learning gains. For instance, linguistic diversity is found to predict intellectual growth, while visual engagement features correlate with social and civic learning. This analysis not only demonstrates the practical relevance of multimodal modelling but also contributes to the broader understanding of learning as a socially embodied process.

1.2 Study Overview

Figure 1.1 illustrates the overall research flow of the thesis. The three studies are designed as successive stages addressing progressively higher levels of human behaviour and interpretation.

The first stage focuses on detecting backchannel behaviours from video sequences of group discussions. This task establishes the methodological foundation by developing motion-based temporal models capable of recognising subtle nonverbal responses across participants.

The second stage extends this framework to infer the meaning of these behaviours by estimating agreement intensity from multimodal inputs. Through audio–visual fusion and role-specific analysis, it captures the temporal evolution of agreement as an attitudinal dimension within interaction.

The final stage applies multimodal analysis to predict learning outcomes in online service-learning programmes. Using linguistic and visual features extracted from authentic discussions, the study connects conversational behaviour with measurable educational outcomes.

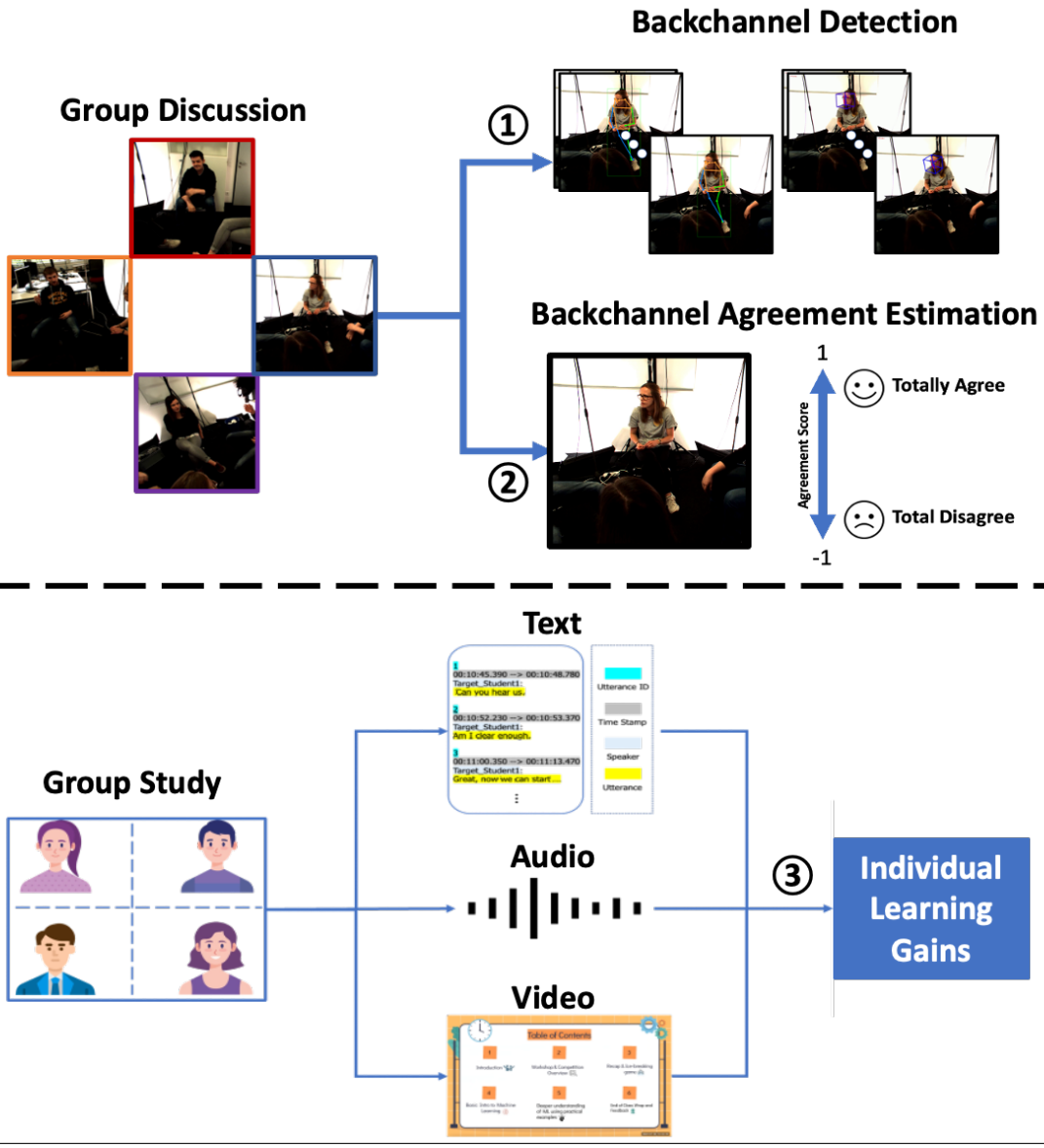
Across all stages, the thesis adopts a multimodal and data-driven approach, integrating computer vision, audio signal processing, and natural language analysis. Each stage builds upon the previous one, moving from behavioural perception to attitudinal inference and, ultimately, to outcome prediction.

1.3 Thesis Aims and Outline

The aims of this thesis are summarised as follows:

- To develop a robust video-based approach for detecting backchannel behaviours in natural group conversations, enabling fine-grained analysis of listener responsiveness and conversational flow.
- To design a multimodal modelling framework for estimating agreement intensity from backchannel behaviours by integrating audio and visual cues with temporal fusion and role-aware processing.
- To investigate how multimodal conversational features relate to students' learning gains in online service-learning programmes and to construct predictive models of learning outcomes across intellectual, social, civic, and intrapersonal domains.

Figure 1.1 Overall conceptual framework of the thesis. The research progresses from (a) video-based backchannel detection, through (b) multimodal agreement estimation, to (c) learning-gain prediction in online service-learning settings.



- To establish a unified computational framework that connects behavioural detection, attitudinal inference, and outcome prediction, advancing multimodal modelling of group interaction.

The remainder of this thesis is organised as follows:

- **Chapter 2: Literature Review** surveys existing work on multimodal group interaction analysis, focusing on backchannel detection, agreement estimation, and learning analytics.
- **Chapter 3: Backchannel Detection in Group Conversation** presents the methodology, experimental setup, and evaluation of the proposed visual models for backchannel detection.
- **Chapter 4: Agreement Estimation from Backchannel Behaviours** introduces the multimodal fusion architecture and temporal modelling framework for estimating agreement intensity.
- **Chapter 5: Predicting Learning Gains in Online Service-Learning Programs** details the modelling of learning outcomes based on multimodal conversational features extracted from authentic educational discussions.
- **Chapter 6: Discussion** discusses the findings across all studies, analyses methodological and conceptual implications, and examines generalisability.
- **Chapter 7: Conclusion and Future Work** summarises the key contributions of the thesis and outlines directions for future research.
- **Chapter 8: Other Contributions** summarises additional publications, collaborative projects, and research outcomes during the Ph.D. study that extend beyond the core scope of this thesis.

Chapter 2

Literature Review

This chapter reviews the theoretical and empirical foundations relevant to understanding and modelling human group interaction. It situates the present work within the broader research on verbal and nonverbal communication, with particular emphasis on listener feedback mechanisms such as backchannels and the expression of agreement. It further examines developments in computational approaches to multimodal interaction analysis and concludes with a discussion of multimodal learning analytics (MMLA) in educational contexts, highlighting the methodological and conceptual framework that supports this thesis.

2.1 Verbal and Nonverbal Communication

Human communication is a rich, multimodal process that involves the coordinated use of verbal and nonverbal signals to exchange information, express emotion, and establish social relationships [48]. Verbal communication operates through language—spoken or written—and provides the structured symbolic system through which people articulate meaning. Words, syntax, and grammar make abstract reasoning and precise reference possible. Spoken communication further incorporates prosodic and paralin-

guistic cues such as intonation, rhythm, pitch, and emphasis, which convey affective nuance, indicate turn boundaries, and shape interpretation [85].

Nonverbal communication, by contrast, encompasses behaviours that operate outside linguistic structure but are deeply integrated into meaning-making. Facial expressions, eye gaze, posture, gestures, and other bodily movements reveal emotional states and relational attitudes often more directly than words [50]. Temporal and spatial factors also contribute to nonverbal expression: proxemics (use of space) and chronemics (timing, latency, and pacing) regulate social distance and engagement [69]. Paralanguage—the non-lexical aspects of voice—serves a similar role, modulating tone, loudness, and tempo to signal emphasis or stance [82].

Rather than functioning independently, these modalities continuously interact. Nonverbal cues complement verbal content by reinforcing meaning or guiding conversational structure—for instance, a nod accompanying "yes" intensifies affirmation, while gaze direction manages turn-taking. They can also contradict verbal messages, such as a sarcastic tone undermining a literal statement. In group contexts, these interdependencies multiply: multiple speakers coordinate overlapping contributions while monitoring others' verbal and visual signals to maintain shared understanding [64, 33].

The integration of verbal and nonverbal modalities therefore forms the basis for social coordination and cognitive alignment in conversation. This interdependence underlies the phenomena examined in this thesis—backchanneling, agreement, and learning-related behaviours—where subtle variations in multimodal cues shape how meaning, attention, and understanding emerge in group interaction.

2.2 Backchannel in Communication

During conversation, listeners play an active role by offering brief, often subtle, feedback known as backchannels. These responses—such as nods, smiles, gaze shifts, or short utterances like "uh-huh," "yeah," or "right"—occur without taking the conversational floor but signal attentiveness and involvement [102, 29]. Backchannels sustain conversational rhythm and help speakers monitor comprehension. In multiparty settings, they also regulate overlapping talk and help distribute attention among participants [12, 18].

Backchannels perform multiple functions. They can indicate understanding, encourage continuation, or convey stance, such as agreement, doubt, or empathy [14]. Scholars often distinguish between *continuers*, which signal attentiveness, and *assessments*, which respond to content [39]. The timing and modality of a backchannel determine its pragmatic meaning. A nod mid-utterance often signals acknowledgment, while a nod following a clause boundary may convey agreement or readiness to take a turn [88].

Both verbal and nonverbal cues contribute to backchanneling. Verbal responses draw from limited lexical forms, while nonverbal feedback—nodding, smiling, eyebrow movements, posture shifts—carries relational and affective nuance [51, 75]. Importantly, backchannels are shaped by culture and context: the frequency, form, and timing vary across languages and social settings [99, 53].

The computational study of backchannels has evolved from descriptive to data-driven approaches. Early models used prosodic rules or pause-based thresholds to predict listener feedback [92, 41]. More recent work employs multimodal machine learning to detect or generate backchannels from acoustic, linguistic, and visual features [63, 71, 2, 17]. Key predictors include head motion, gaze direction, speech energy, and facial expressions, integrated through temporal or transformer-based architectures [95, 83].

Despite advances, challenges persist in modelling backchannels in real-world group settings. Natural interactions involve overlapping speech, variable visibility, and complex attribution—listeners may respond to multiple speakers simultaneously. Accurately identifying to whom a backchannel is directed and interpreting its communicative intent requires temporally precise, multimodal modelling and an understanding of group structure [37, 38]. These challenges motivate the first stage of this thesis, which develops visual-based approaches for backchannel detection in natural group interaction.

2.3 Agreement in Communication

Agreement reflects how interlocutors align cognitively, emotionally, or socially during conversation. It signals shared understanding, approval, or acceptance and helps maintain rapport and conversational flow [77]. In collaborative contexts, agreement supports consensus-building and coordination, while subtle expressions of disagreement indicate cognitive divergence and drive discussion. It is worth noting that agreement as a communicative construct is not monolithic. Scholars have distinguished among several related but distinct acts, including acknowledgment (signalling that a message has been received), endorsement (explicitly supporting a claim), compliance (yielding to social expectation), and affective support (expressing empathy) [88]. These distinctions matter because the same observable behaviour—such as a head nod—may serve different functions depending on conversational context. The computational studies reviewed below, and the work presented in this thesis, primarily model agreement as a graded, composite perception of listener alignment rather than decomposing it into these finer categories; this scope is discussed further in Chapter 4.

Linguistic studies identify lexical and syntactic markers of agreement, such as "yes," "exactly," or "I agree," along with hedges and mitigations that temper disagreement [52, 43]. Beyond language, prosodic and paralinguistic cues—pitch, stress, and

rhythm—modulate intensity and certainty [25, 20]. For example, a rising intonation may express tentative assent, while firm emphasis conveys strong alignment.

Nonverbal cues also play a decisive role. Facial expressions, head nods, smiles, and sustained eye contact all contribute to the perception of agreement or empathy [73, 3]. These cues are not static but unfold dynamically over time. A quick nod during speech may function as acknowledgment, whereas a slower, emphatic nod near a speaker's pause can convey full endorsement [88]. Backchannel behaviours often carry this dual role: while ostensibly signalling attention, they also implicitly indicate alignment or stance toward the ongoing message [19, 77].

Recent computational work has explored agreement estimation using multimodal features. Early approaches focused on lexical or acoustic indicators [98], but the integration of visual signals has significantly improved performance [2, 66]. Deep sequential models such as RNNs, GRUs, and transformers capture temporal dependencies and cross-modal relationships that underlie expressive behaviour [23, 93]. However, the contextual nature of agreement—varying across speakers, cultures, and roles—remains a key challenge for generalisation.

In multi-party conversation, agreement is distributed across individuals and time. Listeners may show alignment through subtle feedback directed at a specific speaker, yet this behaviour also influences group-level perception. Modelling such distributed agreement requires multimodal temporal representations that can capture interpersonal dependencies and contextual variation. This understanding motivates the second stage of this thesis, which focuses on estimating agreement intensity from multimodal backchannel behaviours.

2.4 Multimodal Analytics

Human communication is intrinsically multimodal: meaning emerges from the interplay of language, prosody, facial expressions, gaze, and gesture. Analysing these diverse data streams requires integrative computational approaches capable of capturing temporal and cross-modal relationships [94, 10]. Multimodal analytics offers a framework for studying such complex interactions, combining feature extraction, synchronization, and fusion across heterogeneous modalities.

In human interaction research, multimodal approaches have enabled the automated analysis of turn-taking, engagement, and affective states [36, 13]. Signals from video, audio, and text are transformed into behavioural features—facial action units, gaze direction, speech prosody, and semantic embeddings—which are then modelled to infer social phenomena such as dominance, rapport, or attention [46, 67]. Advances in deep learning have strengthened this capability by learning hierarchical and temporal representations directly from raw multimodal data [10, 54]. Yet challenges persist in synchronisation, interpretability, and scalability, especially for multi-party interaction where behavioural signals overlap and context shifts rapidly.

Multimodal analytics has also gained prominence in educational research, where it informs the study of student engagement, collaboration, and learning processes. Traditional learning analytics primarily rely on log data or assessment outcomes, which provide limited insight into the behavioural and social dimensions of learning [24, 26]. Multimodal Learning Analytics (MMLA) expands this scope by integrating speech, gesture, gaze, and physiological data with traditional educational indicators [70, 60]. Through this integration, researchers can examine how cognitive, emotional, and collaborative processes unfold during authentic learning activities [1, 58].

However, methodological complexity remains a central issue. Combining heterogeneous data streams requires robust feature alignment, temporal segmentation, and context modelling. Furthermore, the interpretability of deep models poses a barrier

to educational deployment, where transparency and pedagogical meaning are essential [65, 61]. Despite these challenges, MMLA holds significant promise for capturing learning as a situated, multimodal process. By linking behavioural cues with outcomes such as engagement or learning gains [34, 78], it provides a framework for understanding how group interaction contributes to individual development.

The studies in this thesis build on these advances. They combine insights from social signal processing and learning analytics to model multimodal group interaction, focusing on backchannel detection, agreement estimation, and the prediction of learning outcomes. Across these three directions, several gaps in the existing literature motivate the present work: (1) most backchannel detection methods have been developed for dyadic settings and do not explicitly model multi-scale temporal dynamics across behavioural channels; (2) agreement estimation has rarely been treated as a continuous, multimodal regression problem with role-aware preprocessing; and (3) multimodal learning analytics in synchronous online education has yet to integrate screen-content and gallery-view modalities alongside linguistic and acoustic features. The following chapters address these gaps through dedicated methodological contributions and empirical evaluations.

Chapter 3

Backchannel Detection in Group Conversation

Backchannel responses constitute an essential part of human communication, enabling listeners to display attention, understanding, and social alignment without interrupting the speaker's flow [42, 90]. These subtle listener cues—such as head nods, gaze shifts, or brief facial expressions—help maintain conversational rhythm and foster mutual awareness. They are so common in everyday dialogue that people rarely notice them, yet conversations without backchannels could be perceived as unnatural or disengaged. Automating their detection is therefore a crucial step toward developing conversational agents that can interpret social signals and respond in a more human-like manner.

Although backchannel behaviours are integral to conversation, detecting them computationally remains a challenging problem. They are brief, vary across individuals and cultural contexts, and are often interwoven with other natural movements such as posture shifts or gaze adjustments. In group conversations, the difficulty increases: multiple participants speak and listen at the same time, producing overlapping gestures and expressions. The same physical movement may convey different meanings

depending on who performs it, when it occurs, and in relation to whom. Reliable detection therefore requires models capable of capturing both temporal dynamics and relational context—that is, how one participant's subtle reactions are shaped by the flow of others' speech and movement. Previous studies have explored verbal and acoustic cues for identifying backchannels, yet these modalities face both practical and cultural constraints. Audio-based features are easily degraded in noisy environments or unavailable in recordings without a dedicated audio channel, while verbal responses differ widely across languages and social norms [42]. Visual cues, by contrast, are continuously observable and relatively universal. They capture spontaneous reactions—such as nodding, smiling, or shifting gaze—that frequently accompany understanding and agreement. This makes them particularly valuable for robust and cross-situational modeling of listener feedback, especially in group settings where multiple streams of nonverbal behaviour coexist.

Guided by this motivation, this chapter investigates visual backchannel detection in multi-party conversation. The goal is to recognise listeners' responsive behaviours from continuous video and to capture how these cues evolve over time across multiple behavioural channels. To this end, we propose the **Temporal Multichannel Attention Network (TMAN)**, a framework designed to model the temporal coordination and interaction among four visual cues: body pose, head orientation, gaze direction, and facial movements. Instead of analysing these cues independently, TMAN learns how their subtle temporal dependencies jointly indicate listener feedback while filtering out incidental motion and background activity. This approach reflects a broader aim of this research—to understand how fine-grained visual behaviours encode social understanding and responsiveness during natural human interaction.

It is worth noting that recent advances in vision transformers and foundation models have achieved impressive results on general video understanding tasks. However, as demonstrated in the experiments reported later in this chapter, these generic architectures—whether used as frozen feature extractors or fine-tuned on the target

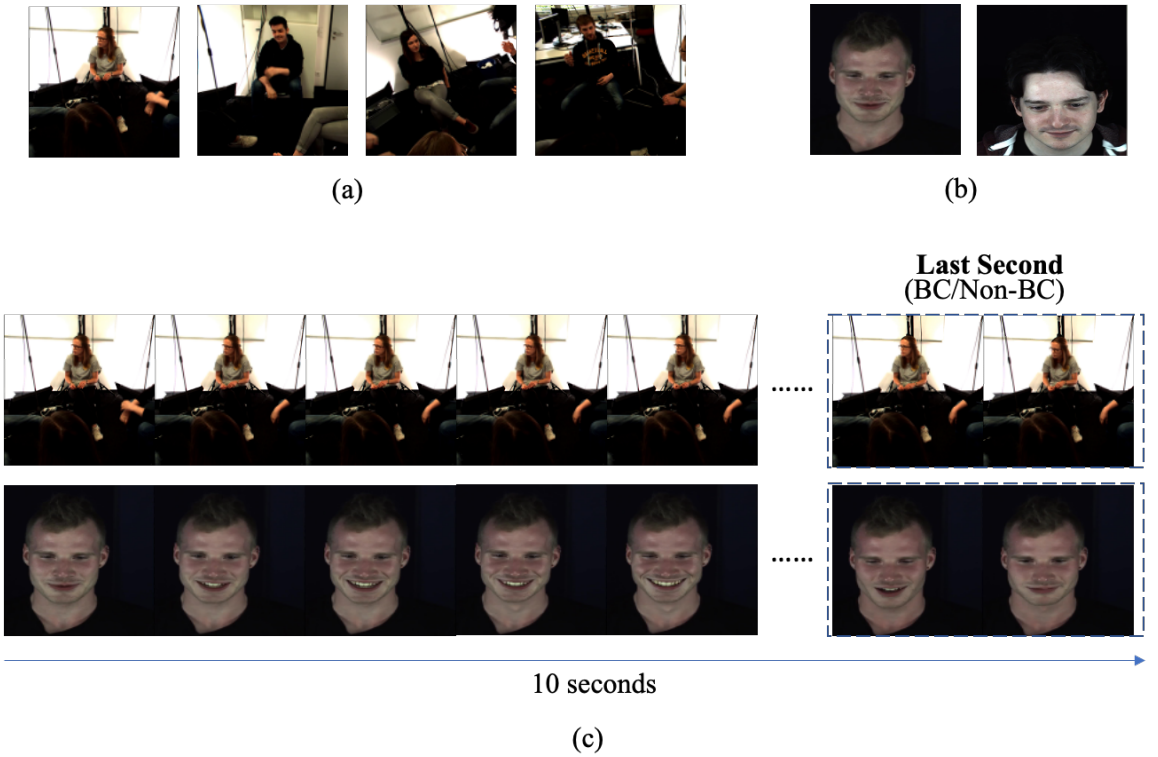
task—do not match the performance of TMAN on backchannel detection. The key distinction lies in TMAN's explicit design for social signal understanding: rather than relying on broad-spectrum visual representations learned from large-scale datasets, TMAN introduces domain-specific inductive biases through motion-aware feature transformations, hierarchical cross-channel attention, and multi-scale temporal reasoning. These design choices reflect the hypothesis that detecting brief, context-dependent social cues in group conversation requires architectural mechanisms tailored to the temporal and relational structure of human interaction, rather than general-purpose visual encoders alone.

The remainder of this chapter is structured as follows. Section 3.1 introduces the datasets used in this study and explains how backchannel instances were constructed from continuous conversational recordings. Section 3.2 details the extraction of visual behavioural features, including body, head, gaze, and facial action-unit cues. Section 3.3 describes the feature transformation strategies designed to capture motion dynamics across time. Section 3.4 presents the proposed Temporal Multichannel Attention Network (TMAN) and its hierarchical attention modules. Section 4.4 reports the experimental results and comparisons with existing approaches, followed by Section 3.6, which analyses the effects of feature engineering, attention mechanisms, and visual-channel combinations. Section 3.7 discusses the key findings and their broader implications. Finally, Section 5.6 summarises the main conclusions of this chapter and outlines directions for subsequent investigation.

3.1 Dataset

To evaluate the proposed framework, this study uses publicly available datasets that capture natural conversational behaviour. However, most existing datasets in human communication research, such as Vyaktitv [45], P2PSTORY [86], and Canal9 [76], are not suitable for automatic backchannel detection. Many, including Vyaktitv and

Figure 3.1 Illustration of the MPIIGroupInteraction and CCDb datasets: (a) example frames from MPIIGroupInteraction showing a full-body view of seated group conversations; (b) example frames from CCDb showing a close-up view (head/face focus) of seated dyadic conversations; and (c) the data instance construction for both datasets. Each row shows frame sequences from a 10-second video segment: the top row corresponds to MPIIGroupInteraction, and the bottom row corresponds to CCDb. The blue dashed boxes highlight the last one-second segment, which is used to determine the video label: "BC" (backchannel) if a backchannel response occurs, or "Non-BC" (non-backchannel) otherwise. Only the target participant's behaviour in the last second is used for label assignment; the preceding frames serve as context.



Canal9, lack explicit annotations of backchannel responses, while others, such as P2PSTORY, label only isolated cues such as head nods or smiles rather than full backchannel events that combine verbal and nonverbal information. These limitations make it difficult to model the multimodal and context-dependent nature of

listener feedback. To address this gap, this chapter employs two public datasets that explicitly annotate backchannel behaviour: *MPIIGroupInteraction* [67, 66] and the *Cardiff Conversation Database (CCDb)* [6]. Both datasets contain video recordings of spontaneous conversation and include multimodal information suitable for analysing visual backchannel behaviour in group and dyadic settings.

MPIIGroupInteraction. The MPIIGroupInteraction dataset includes 22 German group discussions, each about 20 minutes long, adding up to over 440 minutes of recorded interaction. A total of 78 participants (43 female, 35 male), aged between 18 and 38, were recruited from a German university campus, all native German speakers to keep the language consistent. Each group had either three or four people (10 groups of three, 12 groups of four) who sat in a circular arrangement so they could see one another. Several cameras were placed around the setup to record near-frontal views, and individual microphones were used to capture clear audio without blocking faces. Backchannel events were annotated by three trained annotators, each assigned to different parts of the dataset. They followed a standardised protocol based on established definitions of listener feedback. Both auditory and visual signals were considered. Visual annotations included nodding, head shaking, and facial expressions, while auditory annotations captured verbal and paraverbal responses such as "yes," "oh really?," "uh-huh," or "aha." When several cues appeared together, for example nodding while saying "yes," they were marked as a single event. Each participant, on average, showed around 70 backchannel events. Using these annotations, 10-second video segments were created to include both the backchannel moment and its immediate context, with each segment ending at the time of a potential backchannel. Negative clips were selected from stretches without annotated backchannels to keep the data balanced. The dataset used in this study was divided into 6,716 clips for training (3,358 positive and 3,358 negative), 2,854 for validation, and 4,898 clips reserved for official testing, for which the ground-truth labels are not publicly released. The task is thus defined as a binary classification problem to decide whether

a backchannel occurred within each clip. More details on the dataset and annotation procedure are described in [67, 66].

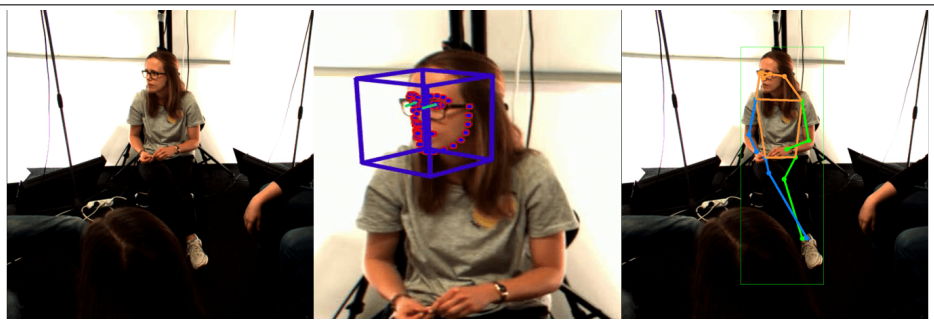
Cardiff Conversation Database (CCDb). The CCDb dataset contains 30 dyadic conversations, each about five minutes long, involving 16 participants (12 male, 4 female) aged 25–56 from Cardiff University in the United Kingdom. In each recording, two speakers sat face-to-face, and both audio and video were captured with a camera–microphone pair placed in front of each person. As shown in Figure 3.1, CCDb shares a similar setup with MPIIGroupInteraction but has closer framing that focuses on facial and head movements.Backchannel annotations were created using the ELAN toolkit [100], following the procedure described by Aubrey et al. [6]. A backchannel was defined as any short facial, gestural, or vocal response triggered by the interlocutor's behaviour, including smiles, nods, and short utterances like "umm" or "err." Following the same segmentation procedure as in MPIIGroupInteraction, 633 backchannel segments were extracted, each 10 seconds long at 30 fps, ending at the time of the labelled event. An equal number of non-backchannel clips were sampled, giving 1,266 clips in total. The data were split by conversation session at a ratio of about 7:3, with 839 clips for training and 427 for validation.

The two datasets capture complementary perspectives on conversation. MPIIGroupInteraction provides a broader view of multi-party discussion with visible body movement and group dynamics, while CCDb focuses on dyadic exchanges that highlight detailed facial and head behaviour. Using both allows for evaluating visual backchannel detection across settings that differ in scale, viewpoint, and social interaction structure.

3.2 Feature Extraction

In this study, we extract a range of nonverbal behavioural features from video to predict listener backchannel responses. Specifically, four behavioural channels are considered—body pose, head pose, eye gaze, and facial action units (AUs)—each capturing a distinct but complementary aspect of listener behaviour. Throughout this chapter, the term *channel* refers to an individual stream of behavioural information extracted from the visual modality. Combined, these cues represent attentional, affective, and interactive patterns that reflect listener feedback in natural conversation. Body pose provides global information about upper-body movement and orientation, which often indicates attentiveness or responsiveness [44]. Head pose captures local motion dynamics such as nodding or shaking, which serve as direct indicators of acknowledgment or agreement [66, 95]. Eye gaze represents shifts in visual attention and turn-taking intention [4], while facial AUs describe subtle muscle activations—such as eyebrow raises or smiles—that convey understanding, agreement, or empathy [4, 21].

Figure 3.2 Visualization of multimodal behavioral features in a video frame. Left: Original video frame of the target individual. Middle: Eye gaze estimation (green vectors), head pose tracking (3D bounding boxes), and facial AUs recognition using OpenFace 2.0. Right: Body pose estimation (human bounding box and body keypoints (e.g., nose and ears)) using ViTPose



As shown in Figure 3.2, two open-source toolkits are used for feature extraction. ViTPose [101], a vision transformer–based pose estimator, is employed to obtain

high-resolution body keypoints, while OpenFace 2.0 [11] extracts facial, head, and gaze-related attributes. These models are selected for their demonstrated accuracy and robustness under varied lighting conditions, partial occlusions, and spontaneous conversational settings. The resulting multichannel visual features serve as the input for temporal modelling in the proposed TMAN framework, and the extraction procedures for each channel are as follows:

Body Pose. Body pose features describe the two-dimensional spatial configuration of human keypoints corresponding to skeletal joints. ViTPose is used to extract (x, y) coordinates for 17 keypoints per frame in the COCO format [57]. In this work, only 11 upper-body keypoints are retained—nose, left and right eyes, left and right ears, left and right shoulders, left and right elbows, and left and right wrists—since these regions provide most of the visible behavioural cues relevant to backchanneling. Prior studies have shown that upper-body movements, including nodding, gesturing, and posture shifts, are strong indicators of active listening and conversational involvement, whereas lower-body motion tends to be minimal in seated discussions and susceptible to occlusion [16, 46, 36]. The extracted joint trajectories form a structured temporal sequence that captures both micro-movements (e.g., brief nods) and sustained gestures.

Head Pose. Head pose features capture the three-dimensional position and orientation of the head relative to the camera. OpenFace detects facial landmarks and estimates six parameters per frame: a 3D position vector (x, y, z) measured in millimetres, and three rotational angles (pitch, yaw, and roll) representing head tilt, turn, and nod. These features characterize dynamic head behaviour, which is among the most salient non-verbal signals of feedback in conversation. For instance, periodic nodding or subtle reorientation toward the speaker often signals understanding or alignment. Head pose sequences therefore serve as a mid-level motion descriptor bridging global body movement and fine-grained facial expressions.

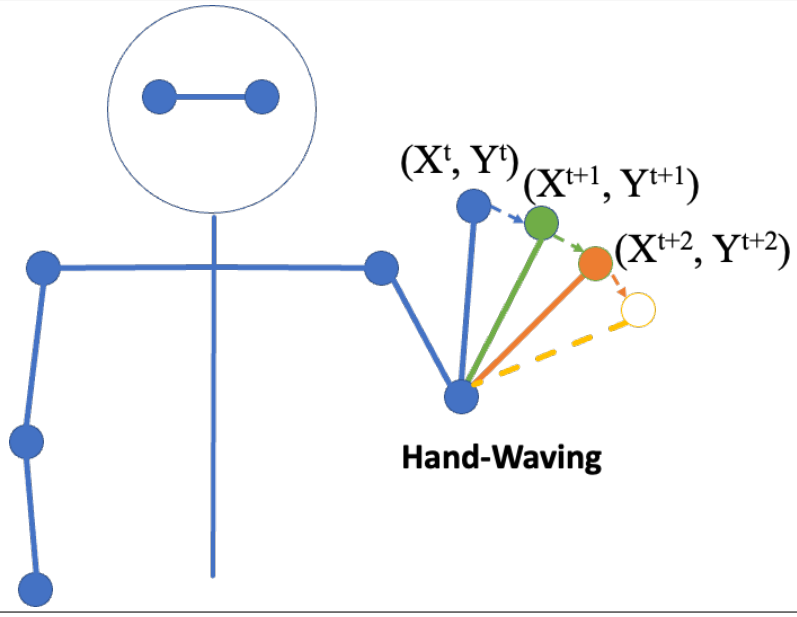
Eye Gaze. Eye gaze features represent the orientation of the listener's visual fo-

cus in three-dimensional space, providing crucial evidence of attention distribution and engagement. OpenFace outputs gaze direction vectors (x, y, z) for both eyes, as well as averaged horizontal and vertical gaze components in radians. The horizontal component reflects whether the listener maintains eye contact or glances away, while the vertical component captures nodding-related changes in gaze trajectory. Gaze patterns not only index attentiveness but also reveal coordination with speaker cues, such as gaze following or turn anticipation, which are highly relevant to backchannel timing.

Facial Action Units (AUs). Facial AUs quantify the movement intensity of specific facial muscles according to the Facial Action Coding System (FACS) [30]. OpenFace extracts the intensity values of 17 AUs, each ranging from 0 (inactive) to 5 (maximum intensity). These include movements such as inner brow raises (AU1), cheek raisers (AU6), lip corner pulls (AU12), and chin raisers (AU17), which together describe affective responses and social alignment. Because the intensity scores already encode the degree of activation, binary presence indicators are excluded to avoid redundancy. Within the context of backchannel detection, AU patterns complement body and head movements by revealing subtle affective cues—such as smiling with a nod—that often accompany agreement or attentive listening.

The integration of these four behavioural channels provides a layered representation of listener behaviour across different spatial and temporal scales. Body and head pose capture broad and mid-level movement patterns, gaze reflects attentional dynamics, and facial AUs convey fine-grained expressive signals. Together, these complementary channels form a coherent basis for modelling visual feedback behaviour. Within the proposed TMAN framework, they enable the network to learn both the temporal progression and the cross-channel relationships that underpin reliable detection of backchannel responses in group conversation.

Figure 3.3 Illustration of skeleton-based motion representation. Consecutive frames show the displacement of the wrist keypoint over time, where each colored joint represents the position at frame t, t+1, t+2, and t+3



3.3 Feature Engineering

The frame-level visual features described above preserve spatial and expressive details of listener behaviour but do not explicitly encode how these cues evolve over time. Backchannel responses, however, are characterised by short, dynamic movements—such as nodding or subtle posture shifts—that unfold over a few consecutive frames. To capture these temporal dynamics, we introduce three feature transformation strategies designed to emphasise inter-frame variation and motion direction.

For each visual channel, temporal differences between consecutive frames are computed as:

$$\Delta_t = \mathbf{f}_{t+1} - \mathbf{f}_t,$$

where $\mathbf{f}_t$ and $\mathbf{f}_{t+1}$ denote the feature vectors at time t and t+1, respectively. Based on this operation, three representations are derived:

1. **Original:** the untransformed frame-level features, which retain static positional or intensity information without highlighting motion;
2. **Abs-Diff:** the absolute difference $|\mathbf{f}_{t+1} - \mathbf{f}_t|$, capturing the magnitude of movement while discarding directionality;
3. **Subtraction:** the signed subtraction $(\mathbf{f}_{t+1} - \mathbf{f}_t)$, which preserves both direction and magnitude of motion, thereby retaining fine-grained temporal information.

Figure 3.3 illustrates this principle using a simplified example of a hand-waving gesture. The coloured wrist joints represent successive positions across frames, forming a motion trajectory across each time step. This directional change carries richer temporal information than static coordinates alone, allowing the model to infer not only the amount of movement but also its evolving trend.

The motion-enhanced features derived from each transformation are organised into four behavioural channels:

- **Body Movement (Body):** displacement of upper-body keypoints across frames, capturing gestures such as nods, arm shifts, and posture adjustments;
- **Head Movement and Rotation (Head):** frame-to-frame changes in head orientation (pitch, yaw, roll) representing nodding, tilting, or turning;
- **Gaze Movement (Gaze):** directional shifts in visual attention, such as gaze redirection or fixation changes toward the speaker;
- **Facial Action Unit Variation (Facial):** inter-frame changes in AU intensity that reveal micro-expressions and fine affective activations.

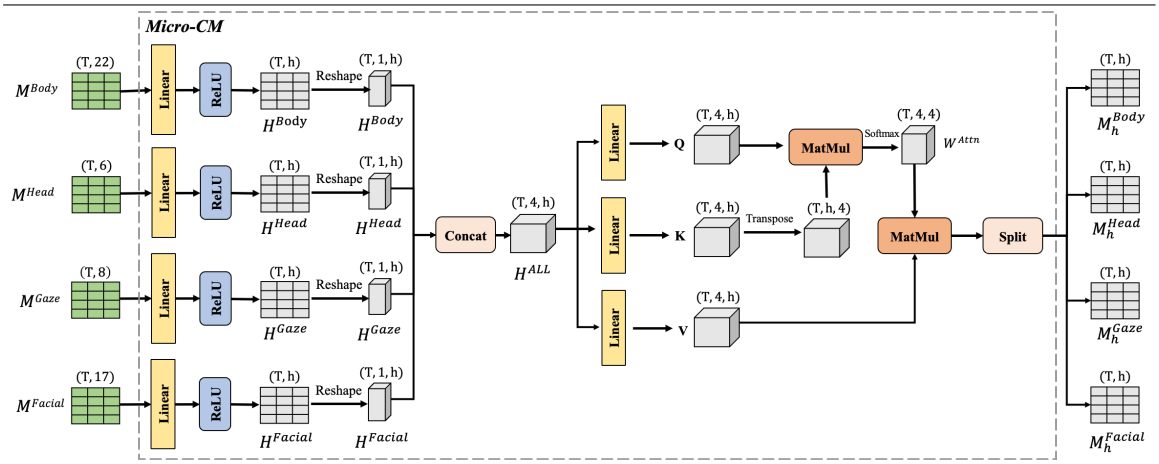
These three representations—*Original*, *Abs-Diff*, and *Subtraction*—enable a systematic investigation of how static, magnitude-based, and directional motion features respectively contribute to recognising the subtle dynamics of listener feedback. The

subtraction-based representation, in particular, provides a more complete description of motion flow and is expected to enhance the model's sensitivity to brief, socially meaningful movements.

3.4 Methodology

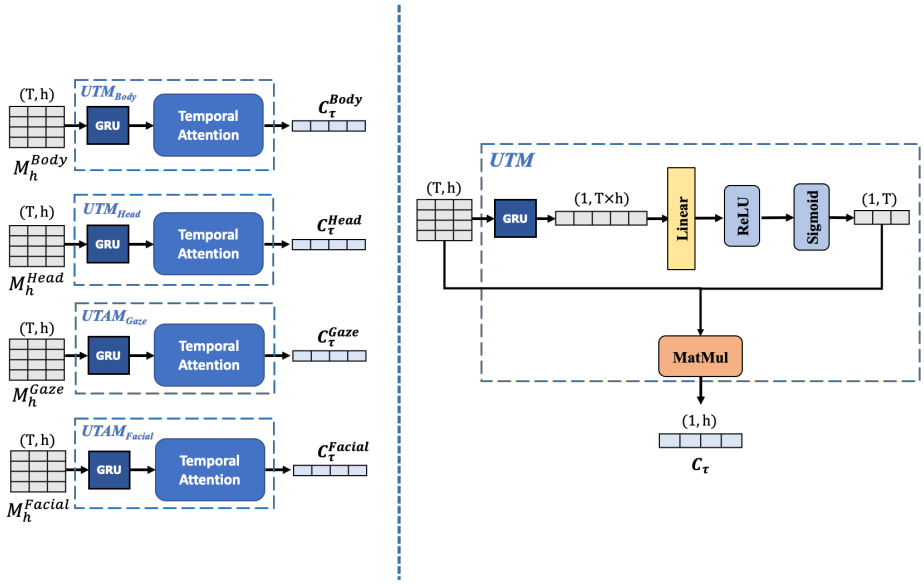
This study proposes the Temporal Multichannel Attention Network (TMAN), a framework developed to address the challenges of visual backchannel detection in group conversation settings. TMAN is designed to capture both the fine-grained temporal dynamics and the interdependencies among multiple behavioural channels extracted from video. The architecture integrates three complementary attention-based components, each responsible for modelling a specific aspect of temporal or cross-channel interaction. Through these mechanisms, TMAN learns to combine cues from body movement, head motion, eye gaze, and facial expressions in an adaptive manner, enabling contextually sensitive and reliable identification of listener backchannel responses.

Figure 3.4 Micro Action Cross-Channel Modelling (Micro-CM) Module



Backchannel signalling is a subtle yet complex part of human communication, involving several behavioural streams that may occur at the same time or in quick succession

Figure 3.5 Unichannel Temporal Modelling (UTM) Module: Left: The overall framework of UTM processing features from four channels. Right: Computation flow of a specific channel within a UTM



within a short observation window. Learning meaningful representations from these diverse visual channels is therefore crucial for accurate backchannel detection. To capture these relationships, TMAN begins with the **Micro Action Cross-Channel Modelling (Micro-CM)** module, which encodes salient visual features while maintaining both temporal coherence and inter-channel relationships. This module uses a self-attention mechanism to evaluate the relative importance of each behavioural channel at each time step, allowing the network to emphasise informative micro actions—such as head nods, gaze shifts, or subtle facial changes—while suppressing irrelevant motion. The resulting representations serve as refined channel features for subsequent temporal processing. An overview of this component is illustrated in Figure 3.4.

Four input feature sequences are provided to Micro-CM, corresponding to *Body Movement*, *Head Movement and Rotation*, *Gaze Movement*, and *Facial Action Unit Vari-*

ation. These are denoted as

$$M^{Body} \in \mathbb{R}^{T\times 22}, \quad M^{Head} \in \mathbb{R}^{T\times 6}, \quad M^{Gaze} \in \mathbb{R}^{T\times 8}, \quad M^{Facial} \in \mathbb{R}^{T\times 17},$$

where T denotes the number of frames within the temporal window.

Each feature stream is first projected into a common latent dimension h using a linear layer followed by a ReLU activation:

$$H^{Body} = \text{ReLU}(W_h^{Body} M^{Body}), \tag{3.1}$$

$$H^{Head} = \text{ReLU}(W_h^{Head} M^{Head}), \tag{3.2}$$

$$H^{Gaze} = \text{ReLU}(W_h^{Gaze} M^{Gaze}), \tag{3.3}$$

$$H^{Facial} = \text{ReLU}(W_h^{Facial} M^{Facial}), \tag{3.4}$$

where W_h^{Body}, W_h^{Head}, W_h^{Gaze}, and W_h^{Facial} are the learnable projection matrices for each channel. The resulting hidden features are reshaped into $(T, 1, h)$ and concatenated along the channel dimension to form a unified sequence:

$$H^{ALL} = \text{Concat}(H^{Body}, H^{Head}, H^{Gaze}, H^{Facial}), \tag{3.5}$$

yielding $H^{ALL} \in \mathbb{R}^{T\times 4\times h}$, which serves as the integrated context representation across all channels.

To capture inter-channel dependencies, the module computes attention weights based on the query (Q), key (K), and value (V) projections of H^{ALL}:

$$Q = W_q H^{ALL}, \tag{3.6}$$

$$K = W_k H^{ALL}, \tag{3.7}$$

$$V = W_v H^{ALL}, \tag{3.8}$$

where W_q, W_k, and W_v are learnable weight matrices. The attention weight matrix is obtained as:

$$W_{Attn} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \tag{3.9}$$

where d_k is the scaling factor equal to the key dimension.

The re-weighted representation is then produced by applying W_{Attn} to V:

$$\tilde{H} = W_{Attn}V.$$

Finally, the output is split back into four channel-specific feature sequences:

$$M_h^{Body},\ M_h^{Head},\ M_h^{Gaze},\ M_h^{Facial} = \text{Split}(\tilde{H}), \tag{3.10}$$

where $M_h^{Body}, M_h^{Head}, M_h^{Gaze}, M_h^{Facial} \in \mathbb{R}^{T\times h}$ denote the refined feature representations for each behavioural channel. Through this process, the Micro-CM module enhances the expressiveness of each visual signal by learning how localised micro actions interact across channels. The refined feature maps are subsequently passed to the temporal encoding stage for higher-level sequence modelling.

The next stage of TMAN focuses on modelling the temporal evolution of each behavioural channel independently. The **Unichannel Temporal Modelling (UTM)** component, illustrated in Figure 3.5, is designed to capture the temporal dependencies that arise within the motion and expression sequences of each visual channel. It operates on the refined feature maps M_h^{Body}, M_h^{Head}, M_h^{Gaze}, and M_h^{Facial} produced by the Micro-CM, which represent the enhanced signals of body movement, head orientation, gaze direction, and facial muscle activation, respectively.

A Gated Recurrent Unit (GRU) network is applied to each feature stream to model how information unfolds over time. GRUs are chosen for their proven efficiency in temporal sequence encoding and their ability to capture long-range dependencies with relatively low computational cost [84]. By encoding the evolution of each modality across frames, UTM learns to represent temporal patterns such as periodic nodding, sustained eye contact, or gradual changes in facial expression—each of which may signal different forms of listener engagement.

Backchannel behaviours can vary considerably in timing and duration, from brief nods to longer expressive gestures. It is therefore important for the model to iden-

tify which moments within a temporal sequence are most indicative of a backchannel response. To achieve this, the UTM integrates an attention mechanism that estimates the saliency of each time step based on its corresponding hidden state from the GRU. The attention weights allow the model to focus more heavily on temporally significant frames while down-weighting those that contribute little to the final representation. This process results in a condensed temporal descriptor that captures the most informative parts of each behavioural sequence.

Formally, the GRU updates for each modality at time step t are defined as:

$$r_t = \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}), \tag{3.11}$$

$$z_t = \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}), \tag{3.12}$$

$$\tilde{h}_t = \tanh(W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{t-1} + b_{hn})), \tag{3.13}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \tag{3.14}$$

where x_t denotes the input feature at time step t, h_t represents the hidden state, and $\sigma(\cdot)$ and $\odot$ indicate the sigmoid activation function and element-wise multiplication, respectively. $W_{ir}, W_{iz}, W_{in}, W_{hr}, W_{hz}, W_{hn}$ and their corresponding biases $b_{ir}, b_{iz}, b_{in}, b_{hr}, b_{hz}, b_{hn}$ are trainable parameters learned during model optimization.

The temporal attention weights a_t are then computed to measure the relative importance of each hidden state:

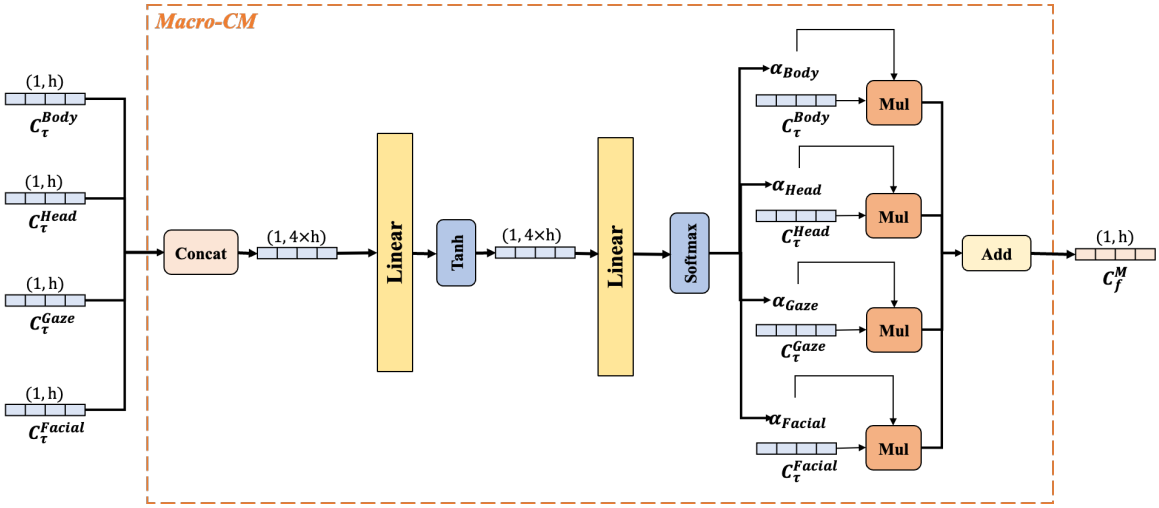
$$a_t = \text{sigmoid}(\text{ReLU}(W_t h_t + b_t)), \tag{3.15}$$

$$c_\tau = \sum_{t=1}^{T} a_t h_t, \tag{3.16}$$

where a_t represents the attention weight at time step t, and $c_\tau \in \mathbb{R}^{1\times h}$ is the context vector that aggregates temporal information across the entire sequence. The parameters W_t and b_t are learnable weights and biases of the attention layer. In this study, the sigmoid function is selected for weighting rather than the conventional softmax, as preliminary experiments showed that it provided more stable attention distributions and better performance for unimodal feature encoding.

This temporal modelling process allows the TMAN framework to retain the essential motion and expression dynamics within each behavioural channel while reducing redundancy. The resulting context vectors c_τ^{Body}, c_τ^{Head}, c_τ^{Gaze}, and c_τ^{Facial} represent temporally summarised unimodal features that are then used in the next stage for cross-channel fusion and final backchannel prediction.

Figure 3.6 Macro Behavior Cross-Channel Modelling (Macro-CM) Module



In the last module, TMAN integrates the temporally encoded representations from all behavioural channels to determine the overall likelihood of a backchannel event. Simply concatenating these features without considering their relative contribution would overlook how different behavioural channels vary in importance across conversational contexts. For instance, a listener might primarily express feedback through head movements such as nodding or shaking, whereas in other cases, backchannel responses may arise from a combination of cues, such as nodding accompanied by smiling or slight posture shifts. To address this, TMAN introduces a **Macro Behaviour Cross-Channel Modelling (Macro-CM)** component, illustrated in Figure 3.6. This component employs an attention mechanism that adaptively estimates the contribution of each channel's temporal representation when forming the final prediction.

The Macro-CM computes attention weights based on the temporal context vectors c_i^τ obtained from the preceding temporal modelling stage, where each vector summarises the sequential dynamics of a specific channel. The attention process is formulated as:

$$a_i = \text{softmax}(\tanh(c_i^\tau W_{m1} + b_{m1})W_{m2} + b_{m2}), \tag{3.17}$$

$$c_f^M = \sum_{i=1}^{N} a_i c_i^\tau, \tag{3.18}$$

where a_i denotes the attention weight assigned to channel i, c_i^τ represents the temporal context vector of that channel, and $c_f^M \in \mathbb{R}^{1\times h}$ is the aggregated multichannel representation. W_{m1} and W_{m2} are learnable weight matrices, while b_{m1} and b_{m2} are their corresponding biases. Through this attention-based fusion, the model learns to dynamically emphasise the most informative channels for a given conversational instance.

Finally, the aggregated multichannel representation is passed through two fully connected layers with a sigmoid activation to predict the probability of backchannel occurrence:

$$\hat{y} = \text{sigmoid}((c_f^M W_{f1} + b_{f1})W_{f2} + b_{f2}), \tag{3.19}$$

where W_{f1}, W_{f2} and b_{f1}, b_{f2} are learnable parameters of the prediction layers, and $\hat{y}$ indicates the final binary prediction—1 for backchannel presence and 0 for absence.

By adaptively weighting the contributions of different behavioural channels, the Macro-CM component ensures that TMAN prioritises the most relevant signals while suppressing less informative ones. This selective integration captures the nuanced interplay of multichannel behavioural cues—such as how gaze alignment reinforces head movement or how facial expressions accompany body gestures—ultimately improving the model's robustness and interpretability in visual backchannel detection.

3.5 Results

This section presents the experimental evaluation of the proposed Temporal Multi-channel Attention Network (TMAN) for visual backchannel detection. We first describe the experimental configuration and training procedure, followed by a detailed analysis of the model's ablation results and comparative performance against state-of-the-art approaches, including large-scale vision foundation models and multimodal large language models (LLMs).

3.5.1 Experimental Setup

Feature extraction follows the procedures described in Section 3.2, using ViTPose and OpenFace 2.0 to obtain motion- and expression-based features. All models are implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU with 24 GB of memory. Each model is trained for 200 epochs with a batch size of 64. The Adam optimizer is used with an initial learning rate of 0.0005 and a weight decay of 0.0001. The hidden dimension h is set to 128 in the first module and 32 in subsequent ones. Within the temporal modelling component, a single-layer GRU with 32 hidden units and a dropout rate of 0.3 is employed. To mitigate overfitting, layer normalization is applied after the GRU layer, and additional dropout layers (with a rate of 0.5) are placed after the attention modules. The binary cross-entropy with logits (BCEWithLogits) loss function is adopted for optimization.

We conduct comprehensive ablation studies to examine the contribution of different architectural choices and preprocessing strategies. Specifically, we test the effect of (1) the three attention-based modules (*Micro-CM*, *UTM*, and *Macro-CM*); (2) various feature transformation methods; (3) different temporal window lengths; (4) recurrent structures; and (5) combinations of visual channels. All ablation studies are performed on the training and validation splits of the MPIIGroupInteraction

and CCDb datasets. Based on these analyses, the optimal configuration—using the *Subtraction* transformation with a 3-second observation window—is selected for final testing and comparison with existing methods.

3.5.2 Evaluation of the TMAN Framework

To validate the effectiveness of the proposed TMAN architecture, we evaluate its components and their combinations under consistent settings. The full model integrates all three attention-based components—*Micro-CM* for cross-channel micro-action encoding, *UTM* for unimodal temporal modelling, and *Macro-CM* for adaptive multichannel integration. Models employing subsets of these modules are compared to determine their contribution to overall performance. In addition, we evaluate different time-window sizes and feature transformation methods to identify optimal temporal granularity and motion representation. The results of these analyses demonstrate that the complete TMAN architecture achieves the most stable and accurate detection performance across both datasets, confirming that each module contributes distinct and complementary strengths in capturing multiscale behavioural cues.

After determining the best configuration, we train the full TMAN model on the complete training data and evaluate it on the held-out test set. Table 3.1 summarises the results in comparison with published benchmarks. TMAN achieves a test accuracy of 70.1%, outperforming existing methods such as Wang et al. [95] (68.6%), Amer et al. [2] (66.4%), and Sharma et al. [83] (62.1%). The model also substantially exceeds the baseline systems from Müller et al. [66], which achieved accuracies between 50–60%. The superior performance of TMAN reflects its ability to jointly model micro-level dynamics, temporal dependencies, and cross-channel interactions—an aspect not fully addressed by earlier approaches that focus on single channels or lack adaptive attention mechanisms.

Table 3.1: Comparison of test set performances with existing state-of-the-art models

Method	Test Accuracy
TMAN (Ours)	**0.701**
Wang et al. [95]	0.686
Amer et al. [2]	0.664
Anonymous 1	0.658
Ma et al.	0.656
Sharma et al. [83]	0.621
Baseline '22 (Head + Pose) [66]	0.596
Baseline '22 (All) [66]	0.592
Baseline '22 (Trivial) [66]	0.500

3.5.3 Comparison with Vision Foundation Models and Multimodal LLMs

To further situate the performance of TMAN, we benchmark it against several categories of recent vision and multimodal models, evaluated on both the MPIIGroupInteraction and CCDb datasets (Table 3.2). These comparisons highlight how TMAN differs from generic pre-trained architectures and task-agnostic large models in handling social behavioural cues.

- **Frozen foundation model feature extractors:** Vision-language foundation models such as CLIP [79] and BLIP2 [55] are used as frozen feature encoders. Their pre-trained embeddings are combined with a GRU-based temporal decoder to perform backchannel classification. This setup examines how well large-scale representations transfer to conversational behaviour understanding without domain-specific adaptation.

- **Fine-tuned multimodal transformers:** Transformer-based video models in-

cluding X-CLIP [59], VideoMAE [91], VideoMAE-V2 [97], and ViViT [5] are fine-tuned directly on the backchannel detection datasets. These models can adapt high-level visual and motion representations to the specific downstream task through supervised training on the training videos.

- **Zero-shot multimodal LLMs:** Instruction-tuned multimodal large language models such as QwenVL2.5 [8] and Video-LLaMA3 [103] are evaluated in a zero-shot setting. Without additional fine-tuning, they rely on generalised cross-modal reasoning to interpret visual and behavioural cues.

Table 3.2: Comparison of different vision foundation models and multimodal LLMs on MPIIGroupInteraction and CCDb datasets.

Approach	Model	MPIIGroupInteraction	CCDb
Frozen Model	CLIP [79]	54.0%	64.3%
Feature Extractors	BLIP2 [55]	60.3%	66.5%
Fine-tuned Multimodal Transformers	X-CLIP [59]	50.0%	56.5%
	VideoMAE [91]	58.1%	71.8%
	VideoMAE-V2 [97]	68.3%	83.5%
	ViViT [5]	59.6%	70.9%
Zero-shot Multimodal LLMs	QwenVL2.5 [8]	51.5%	55.5%
	Video-LLaMA3 [103]	51.3%	50.4%
TMAN (Ours)	—	**76.7%**	**87.8%**

On the MPIIGroupInteraction dataset, fine-tuned video transformers outperform frozen feature extractors and zero-shot multimodal LLMs by a large margin. Among them, VideoMAE-V2 reaches 68.3% accuracy, reflecting the benefit of adapting large-scale video representations to the specific behavioural domain. However, even this strong baseline falls short of TMAN's 76.7%. The difference highlights a key lim-

itation of generic video transformers: while they capture broad motion dynamics, they are not optimised for the micro-temporal coordination and subtle expressiveness that characterise backchannel responses. In contrast, TMAN explicitly models fine-grained motion signals (e.g., nodding, micro-expressions) and integrates them through temporal and cross-channel attention, enabling it to discern meaningful listener cues from background movement. A similar pattern emerges on the CCDb dataset, where TMAN achieves 87.8% accuracy, surpassing all baseline models including VideoMAE-V2 (83.5%). The overall higher accuracy across methods on CCDb is likely due to its dyadic setup, which produces clearer frontal views and fewer occlusions compared to the more visually complex group scenes in MPIIGroupInteraction. Yet even in this relatively simpler setting, foundation and zero-shot models lag behind, showing that high-level generalisation alone is insufficient for interpreting nuanced social feedback behaviours.

These findings reveal several important insights. First, large pre-trained visual models excel at recognising general visual patterns but are not inherently sensitive to short, transient cues that depend on interaction context. Second, fine-tuning improves representation alignment but still overlooks the need for multi-scale temporal reasoning across behavioural channels. Finally, TMAN's architecture—built around motion-aware feature transformations and hierarchical attention—demonstrates that explicit temporal modelling and channel-level integration remain essential for decoding social signals that unfold over time. This combination allows TMAN to bridge the gap between perceptual understanding and social interpretation, achieving both higher accuracy and better interpretability in backchannel detection.

Table 3.3: Ablation study of feature transformations and observation windows on the MPIIGroupInteraction dataset. The underlined value denotes the best validation accuracy within each model and transformation setting

Model	Feature Transformation	Validation Accuracy				
		1 Sec.	2 Sec.	3 Sec.	4 Sec.	5 Sec.
Micro-CM	Ori. Value	59.8%	**60.2%**	59.9%	59.5%	59.3%
	Abs. Diff	67.2%	69.4%	**71.4%**	71.1%	70.2%
	Subtraction	70.3%	71.1%	**73.5%**	72.7%	72.4%
UTM	Ori. Value	58.5%	59.4%	**60.2%**	59.2%	58.8%
	Abs. Diff	68.2%	70.1%	**72.5%**	71.3%	71.5%
	Subtraction	69.3%	72.5%	**74.1%**	73.4%	72.2%
Macro-CM	Ori. Value	58.6%	58.8%	**59.6%**	58.8%	58.7%
	Abs. Diff	69.5%	71.4%	72.1%	**72.8%**	72.3%
	Subtraction	71.1%	73.5%	**74.8%**	72.9%	72.0%
Micro-CM+UTM	Ori. Value	58.4%	**59.0%**	58.2%	58.1%	58.2%
	Abs. Diff	69.2%	70.8%	**72.6%**	72.1%	71.9%
	Subtraction	72.1%	73.9%	**75.5%**	75.3%	75.1%
UTM+Macro-CM	Ori. Value	60.3%	62.1%	**63.4%**	63.2%	62.8%
	Abs. Diff	69.7%	70.5%	**71.2%**	**71.2%**	70.9%
	Subtraction	72.1%	73.8%	**75.2%**	74.4%	74.1%
Micro-CM+Macro-CM	Ori. Value	58.5%	59.5%	**59.6%**	59.2%	59.1%
	Abs. Diff	68.8%	70.6%	**71.3%**	71.1%	70.4%
	Subtraction	70.8%	74.4%	**75.4%**	74.5%	74.5%
TMAN	Ori. Value	58.8%	60.7%	**61.3%**	61.0%	60.9%
	Abs. Diff	71.2%	72.7%	**73.5%**	73.3%	73.1%
	Subtraction	74.1%	75.2%	**76.7%**	75.9%	75.6%

Table 3.4: Ablation study of feature transformations and observation windows on the CCDb dataset. The underlined value denotes the best validation accuracy within each model and transformation setting

Model	Feature Transformation	Validation Accuracy				
		1 Sec.	2 Sec.	3 Sec.	4 Sec.	5 Sec.
Micro-CM	Ori. Value	70.3%	70.0%	71.9%	71.4%	**72.1%**
	Abs. Diff	77.5%	82.9%	84.8%	85.5%	**85.9%**
	Subtraction	71.7%	77.8%	82.4%	85.5%	**86.1%**
UTM	Ori. Value	70.5%	70.0%	75.4%	73.5%	**77.3%**
	Abs. Diff	76.1%	80.8%	82.2%	83.4%	**84.8%**
	Subtraction	75.9%	78.2%	80.8%	84.3%	**86.2%**
Macro-CM	Ori. Value	71.2%	**71.7%**	71.4%	71.0%	71.0%
	Abs. Diff	76.1%	79.9%	83.4%	**83.8%**	83.6%
	Subtraction	68.9%	71.7%	79.6%	80.1%	**80.6%**
Micro-CM+UTM	Ori. Value	68.9%	68.4%	69.1%	71.0%	**71.2%**
	Abs. Diff	76.8%	81.7%	83.1%	82.9%	**84.8%**
	Subtraction	74.7%	82.7%	**85.7%**	85.5%	84.5%
UTM+Macro-CM	Ori. Value	66.3%	68.6%	73.8%	71.2%	**74.5%**
	Abs. Diff	76.3%	81.5%	82.4%	83.6%	**84.5%**
	Subtraction	78.7%	81.0%	82.0%	84.8%	**85.5%**
Micro-CM+Macro-CM	Ori. Value	**69.1%**	68.9%	**69.1%**	67.0%	67.9%
	Abs. Diff	77.0%	82.4%	85.0%	85.0%	**85.6%**
	Subtraction	71.2%	79.4%	82.7%	84.1%	**85.7%**
TMAN	Ori. Value	69.8%	67.0%	70.0%	70.3%	**71.4%**
	Abs. Diff	77.5%	84.5%	**85.0%**	84.8%	**85.0%**
	Subtraction	75.2%	83.8%	84.3%	83.6%	**87.8%**

3.6 Ablation Studies

3.6.1 Observation Windows and Feature Engineering

The selection of an observation window and the design of feature representations are central to effective feature engineering for backchannel detection. The observation window defines the amount of temporal context available to the model before it predicts a listener response, while the feature representation determines how motion and appearance information within that context are encoded for learning. Both factors directly influence how well the model captures the timing, dynamics, and responsiveness characteristic of conversational coordination. Tables 3.3 and 3.4 present the validation accuracies obtained from different models across observation windows ranging from one to five seconds, under three feature representation strategies: Original Value, Absolute Difference (*Abs-Diff*), and Subtraction. Two clear trends emerge. First, model accuracy generally improves as the observation window increases from one to three seconds, after which further extension yields diminishing or slightly negative returns. Second, the subtraction-based representation consistently delivers the highest accuracy across nearly all configurations.

On the **MPIIGroupInteraction** dataset, most models reach their peak performance at a three-second window. Micro-CM improves from 59.8% at one second to 73.5% at three seconds, UTM from 58.5% to 74.1%, and Macro-CM from 58.6% to 74.8%. Combined configurations show similar behaviour: Micro-CM+UTM achieves 75.5%, UTM+Macro-CM 75.2%, and Micro-CM+Macro-CM 75.4%. The full TMAN framework attains the highest validation accuracy of 76.7% when using the subtraction-based representation with a three-second window. Beyond this duration, accuracy either plateaus or declines slightly, suggesting that extending the window introduces redundant or weakly related information. These results indicate that for MPIIGroupInteraction, where conversational exchanges are rapid and densely interactive, a three-

second span provides sufficient context to capture the speaker's cues and the listener's immediate feedback without incorporating unrelated segments of dialogue. In contrast, the **CCDb** dataset exhibits a different temporal pattern. Validation accuracy continues to rise with longer observation windows, with most models performing best at five seconds. Using the subtraction-based representation, Micro-CM increases from 71.7% at one second to 86.1% at five seconds, UTM from 75.9% to 86.2%, and Macro-CM from 68.9% to 80.6%. Combined configurations also improve progressively, with Micro-CM+UTM and Micro-CM+Macro-CM both reaching 85.7%, and UTM+Macro-CM achieving 85.5%. The full TMAN model obtains the highest validation accuracy of 87.8% at a five-second window. The longer temporal span proves beneficial for CCDb because its dyadic interactions unfold at a slower conversational pace. Listener feedback often develops gradually through sustained head movement, gaze adjustment, or posture change, and a longer window allows the model to capture these continuous behavioural progressions rather than treating them as isolated gestures.

Across both datasets, subtraction-based representations consistently outperform the other two methods. On MPIIGroupInteraction, subtraction improves accuracy by approximately 2–3 percentage points over *Abs-Diff* and by more than 10 points compared with unprocessed features. In CCDb, the improvement reaches up to 12 percentage points relative to static representations. This advantage arises because subtraction highlights inter-frame differences that emphasise motion and temporal variation while reducing static or background information. Since backchannel behaviours are inherently dynamic—manifested through short, temporally structured movements—representations that encode directional change enable the model to capture these fine-grained variations more effectively. Overall, these findings demonstrate that optimal feature engineering for backchannel detection requires balancing temporal coverage and motion sensitivity. The ideal observation window depends on the interactional rhythm of the dataset: shorter windows better capture the fast and re-

active exchanges typical of group discussions, whereas longer windows suit the slower and more deliberate dynamics of dyadic conversation. Regardless of dataset, motion-sensitive representations derived from frame subtraction enhance model robustness by focusing learning on the subtle behavioural transitions that convey listener engagement.

3.6.2 Attention Modules

A detailed ablation study was conducted to examine the contribution of each attention module—Micro Action Cross-Channel Modelling (Micro-CM), Unichannel Temporal Modelling (UTM), and Macro Behaviour Cross-Channel Modelling (Macro-CM)—as well as their various combinations within the TMAN framework. The goal was to evaluate how local cross-channel encoding, unimodal temporal dynamics, and global multimodal integration collectively influence model performance across different feature representations and observation windows.

Across both datasets, every individual module improves the baseline performance, confirming that each one captures a distinct level of behavioural information. On the **MPIIGroupInteraction** dataset (Table 3.3), Micro-CM, UTM, and Macro-CM achieve 73.5%, 74.1%, and 74.8% validation accuracy, respectively, when using the subtraction-based representation with a three-second observation window. When any two modules are combined, accuracy increases further: Micro-CM+UTM reaches 75.5%, UTM+Macro-CM achieves 75.2%, and Micro-CM+Macro-CM attains 75.4%. The full TMAN model, which integrates all three modules, achieves the highest validation accuracy of 76.7%. This progression demonstrates that the modules provide complementary functions—Micro-CM captures short-term inter-channel relations, UTM models intra-channel temporal dependencies, and Macro-CM fuses these temporal signals into a coherent multimodal representation. Results on the **CCDb** dataset (Table 3.4) show a similar pattern, though with overall higher accuracy due to the

clearer visual framing and simpler dyadic setting. Under the same subtraction-based representation and a five-second window, Micro-CM, UTM, and Macro-CM achieve 86.1%, 86.2%, and 80.6%, respectively. The pairwise combinations again provide performance gains, with Micro-CM+UTM and Micro-CM+Macro-CM both reaching 85.7%, and UTM+Macro-CM achieving 85.5%. The full TMAN attains the highest accuracy of 87.8%. These consistent improvements across two datasets with differing conversational dynamics indicate that the three modules contribute distinct but complementary forms of information integration.

Functionally, the three modules operate hierarchically. Micro-CM enhances the expressiveness of each frame by identifying the cross-channel coordination of localised micro actions, such as nodding accompanied by gaze shifts or subtle facial activations. UTM then models the temporal progression of each channel independently, learning how motion or expression patterns evolve over time within the observation window. Macro-CM integrates the temporally summarised representations and adaptively assigns weights to each behavioural channel based on their contextual relevance. This hierarchy allows TMAN to represent both fast reactive movements and slower, sustained cues within a unified temporal framework. The ablation analysis shows that the integration of Micro-CM, UTM, and Macro-CM is not merely additive but synergistic. Each module contributes a distinct analytical perspective—local, temporal, and global—and their combination results in a more discriminative and robust understanding of listener behaviour. By hierarchically aligning motion, temporal dynamics, and cross-channel interactions, TMAN effectively captures the behavioural patterns that underlie backchannel signalling in group conversation.

3.6.3 Visual features

We systematically evaluate different combinations of the four visual features—body movement, head movement and rotation, gaze movement, and facial action units

Table 3.5: Evaluation of the impact of different modality combinations on TMAN on MPIIGroupInteraction dataset

Modality				Validation Accuracy				
Body Move.	Head Move. & Rota.	Gaze Move.	Facial AUs Var.	1 Sec.	2 Sec.	3 Sec.	4 Sec.	5 Sec.
✓				50.3%	50.6%	**51.1%**	50.6%	50.8%
	✓			70.2%	73.2%	**74.4%**	74.0%	73.9%
		✓		51.8%	53.7%	**55.7%**	54.8%	54.3%
			✓	59.7%	62.9%	63.5%	**64.1%**	63.1%
✓	✓			70.4%	73.9%	**74.7%**	74.0%	73.9%
✓		✓		51.1%	52.1%	**54.2%**	51.9%	51.5%
✓			✓	59.7%	62.8%	**63.8%**	62.7%	62.6%
	✓	✓		70.5%	73.9%	**75.1%**	74.0%	73.5%
	✓		✓	71.7%	**75.7%**	75.4%	74.8%	74.2%
		✓	✓	59.8%	62.9%	**64.1%**	63.9%	63.0%
✓	✓	✓		70.6%	74.2%	**74.6%**	74.2%	73.6%
✓	✓		✓	71.8%	**75.6%**	75.1%	74.8%	73.9%
✓		✓	✓	59.4%	62.4%	**63.6%**	63.5%	63.1%
	✓	✓	✓	71.5%	74.9%	**75.4%**	74.6%	73.8%
✓	✓	✓	✓	72.0%	75.1%	**76.7%**	75.3%	74.9%

Table 3.6: Evaluation of the impact of different modality combinations on TMAN on CCDb dataset

Modality				Validation Accuracy				
Body Move.	Head Move. & Rota.	Gaze Move.	Facial AUs Var.	1 Sec.	2 Sec.	3 Sec.	4 Sec.	5 Sec.
✓				45.0%	52.9%	53.2%	**57.6%**	45.0%
	✓			63.2%	74.0%	77.3%	77.3%	**79.4%**
		✓		54.3%	52.1%	54.8%	**55.0%**	**55.0%**
			✓	73.1%	74.9%	75.4%	77.3%	**77.8%**
✓	✓			71.0%	82.7%	**84.3%**	83.1%	83.6%
✓		✓		45.0%	58.1%	56.9%	60.7%	**65.3%**
✓			✓	71.0%	76.8%	79.4%	83.1%	**84.5%**
	✓	✓		71.2%	81.5%	81.7%	80.8%	**84.3%**
	✓		✓	75.2%	75.2%	78.2%	83.4%	**84.6%**
		✓	✓	74.0%	78.5%	80.3%	82.7%	**84.1%**
✓	✓	✓		70.7%	82.4%	**83.4%**	81.7%	79.4%
✓	✓		✓	72.6%	83.8%	83.4%	82.9%	**84.3%**
✓		✓	✓	73.5%	81.5%	79.9%	82.9%	**83.4%**
	✓	✓	✓	73.3%	81.0%	82.7%	82.0%	**84.8%**
✓	✓	✓	✓	74.0%	76.6%	79.9%	83.6%	**87.8%**

(AUs) variation—using the optimal TMAN configuration with the subtraction-based feature representation. The validation results are summarised in Table 3.5 for the MPIIGroupInteraction dataset and Table 3.6 for the CCDb dataset. Each row in the tables corresponds to a unique combination of input channels, where the checkmarks indicate which behavioural channels are included in the model.

For the **MPIIGroupInteraction** dataset, head movement and rotation alone provide the highest accuracy among single-channel inputs, reaching 74.4% with a three-second observation window. Facial AUs variation also contributes strongly, achieving 64.1% at four seconds, while body and gaze movement yield lower accuracies of 51.1% and 55.7%, respectively. These results indicate that local head dynamics are the most reliable single-channel indicators of listener feedback in group interaction, reflecting their dominant role in conveying acknowledgment and attentiveness. A similar trend is observed for the **CCDb** dataset. Head movement and rotation again achieve the highest performance as a single channel, reaching 79.4% at a five-second window. Facial AUs follow closely at 77.8%, while body and gaze movement remain less discriminative (57.6% and 55.0%, respectively). Although performance for all channels improves slightly as the observation window increases, head and facial cues remain the most informative across window lengths.

Integrating multiple channels consistently enhances detection accuracy across both datasets. In the MPIIGroupInteraction dataset, combinations that include head and facial channels—such as head+facial or body+head+facial—achieve strong performance, while the full TMAN incorporating all four channels attains the highest accuracy of 76.7% with a three-second window. On the CCDb dataset, where interactions are slower and facial details are more prominent, the complete four-channel model achieves 87.8% accuracy with a five-second window. These results demonstrate that individual behavioural channels each provide partial evidence of listener feedback, but their integration yields a more complete and reliable representation of backchannel behaviour. Head and facial channels capture immediate expressive responses, whereas

body and gaze channels contribute contextual cues related to posture and attention. The combined use of all four channels enables the model to interpret the subtle coordination among these behavioural signals, reflecting the inherently multimodal nature of human social communication.

Figure 3.7 Visualization of attention mechanisms during an example inference from the MPIIGroupInteraction dataset. (a) The last three-second frame sequence of the target listener showing a nodding backchannel response. (b) Cross-channel attention from the *Micro-CM* module emphasizing increased focus on head movement. (c) Temporal attention from the *UTM* module aligning with the period of head nodding. (d) Attention weights from the *Macro-CM* module showing that the head movement channel is the most dominant

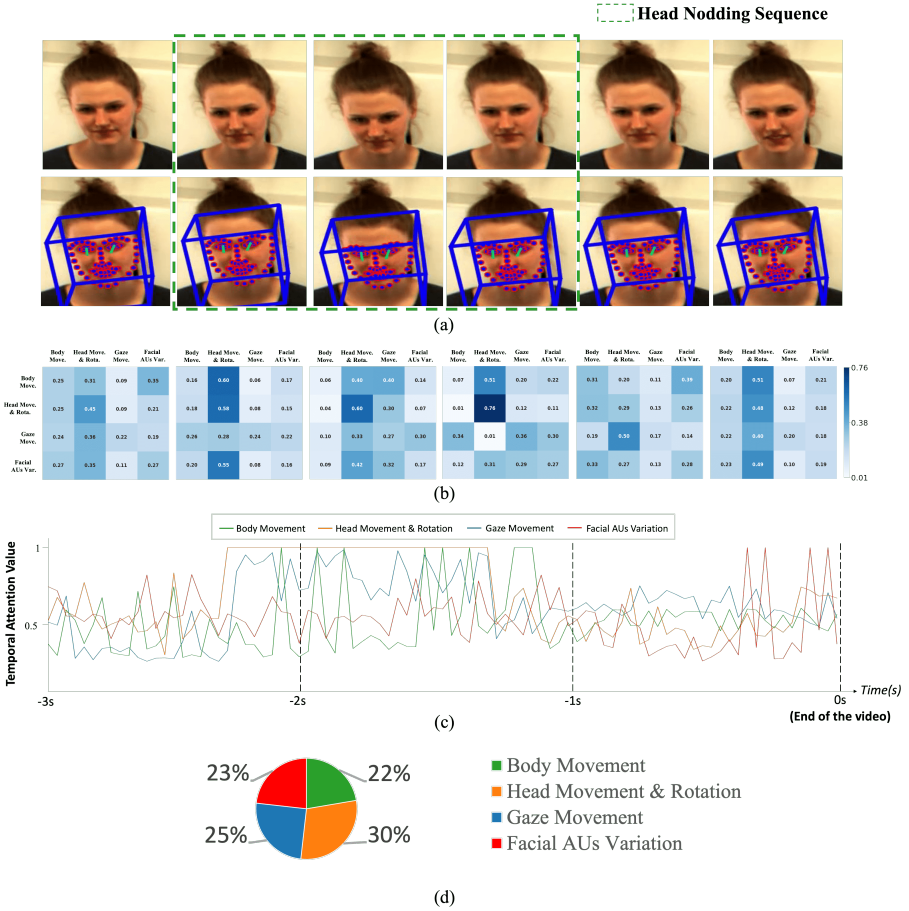
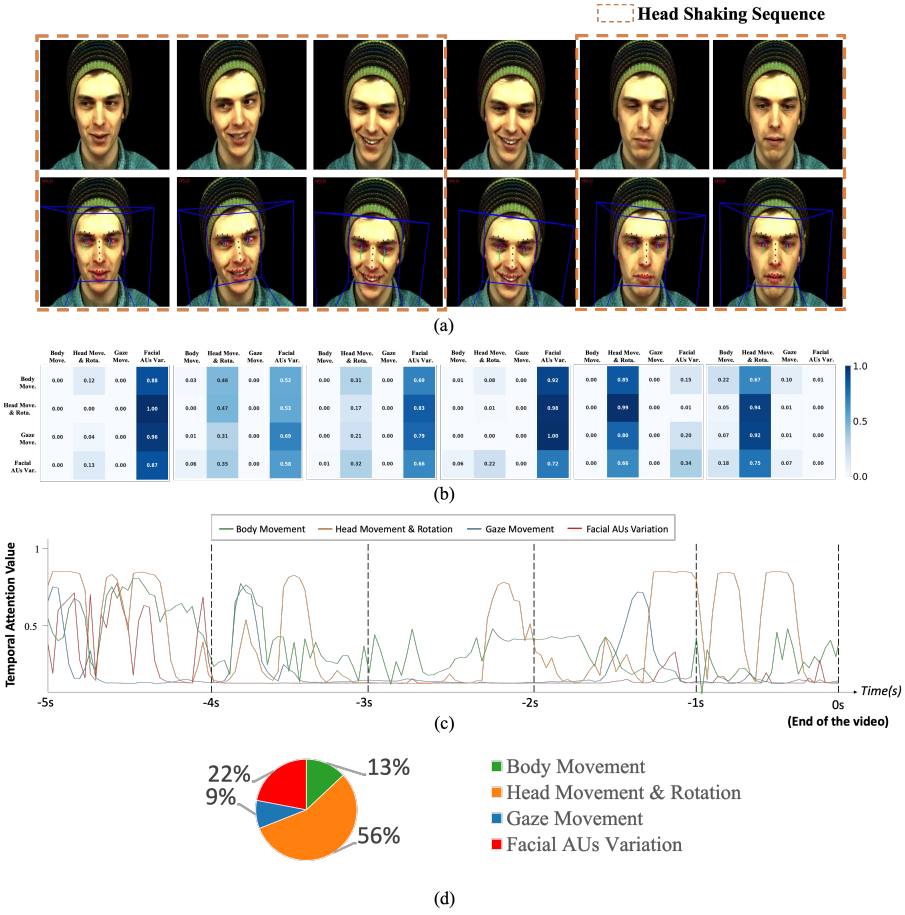


Figure 3.8 Visualization of attention mechanisms during an example inference from the CCDb dataset. (a) The last five-second frame sequence of the target listener showing a head-shake backchannel response. (b) Cross-channel attention from the *Micro-CM* module emphasizing increased focus on head movement and facial action-unit variation. (c) Temporal attention from the *UTM* module aligning with the period of the head-shake response. (d) Attention weights from the *Macro-CM* module showing that the head and facial channels are relatively important



3.7 Discussion

Impact of Observation Window and Feature Engineering Methods. The findings of this study underscore the joint importance of temporal context and fea-

ture transformation strategies in shaping model performance for backchannel detection. Both dimensions play complementary roles: feature transformation governs how salient motion patterns are represented, while observation window length determines how much temporal context the model can effectively utilize. Among the examined transformation techniques, the motion-focused *Subtraction* approach consistently yields the highest performance across models and datasets. By encoding frame-to-frame directional differences, this method accentuates fine-grained temporal variations—such as head nods, gaze shifts, and subtle posture adjustments—that are emblematic of nonverbal listener feedback. In contrast, the *Abs-Diff* transformation captures only the magnitude of motion changes, discarding directional information that is often essential for characterizing rhythmic and asymmetric behaviors inherent in backchanneling. The superior performance of the *Subtraction* transformation thus highlights the value of explicitly modelling directional dynamics to capture the temporal evolution of nonverbal cues.

Temporal window size further mediates the balance between contextual richness and noise accumulation. A short window may fail to capture sufficient behavioral precursors, while an overly long one risks diluting relevant motion patterns with unrelated or redundant activity. On the MPIIGroupInteraction dataset, a three-second observation window achieves the optimal balance, providing enough temporal coverage to capture pre-response cues while avoiding excessive contextual drift. In contrast, the CCDb dataset exhibits peak performance at a five-second window, particularly when paired with the *Subtraction* transformation. This suggests that the temporal extent of meaningful behavioral dependencies may vary with interaction type and dataset characteristics. The relatively longer optimal window on CCDb implies that dyadic, close-up conversations may involve slower or more extended gestural rhythms compared with the faster-paced group discussions in MPIIGroupInteraction.

To further substantiate these observations, a McNemar's test [62] was conducted to examine whether the observed improvements across preprocessing methods were

Table 3.7: Statistical significance (McNemar's test) of different observation windows and feature engineering methods

Observation Window	**Comparison**	**MPII**		**CCDb**	
		χ^2	p	χ^2	p
1 sec.	Subtraction vs Ori. Value	252.546	**0.000**	7.206	**0.007**
	Subtract vs Abs. Diff.	10.190	**0.001**	0.000	1.000
2 sec.	Subtraction vs Ori. Value	346.037	**0.000**	41.059	**0.000**
	Subtract vs Abs. Diff.	12.187	**0.000**	0.583	0.445
3 sec.	Subtraction vs Ori. Value	320.350	**0.000**	45.581	**0.000**
	Subtract vs Abs. Diff.	9.515	**0.002**	4.661	**0.031**
4 sec.	Subtraction vs Ori. Value	220.952	**0.000**	75.155	**0.000**
	Subtract vs Abs. Diff.	9.440	**0.002**	4.167	**0.041**
5 sec.	Subtraction vs Ori. Value	260.269	**0.000**	55.855	**0.000**
	Subtract vs Abs. Diff.	1.309	0.253	4.321	**0.038**

statistically significant. The McNemar test evaluates paired classification results by comparing the number of discordant predictions between two models, using the chi-square statistic:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}, \tag{3.20}$$

where b and c represent the counts of samples misclassified by one model but correctly classified by the other. This non-parametric test is particularly suited for verifying differences in paired binary outcomes, independent of distributional assumptions. As reported in Table 3.7, the results confirm that the *Subtraction* preprocessing method yields statistically significant improvements over the *Original Value* and *Abs-Diff* representations across nearly all observation windows on both datasets ($p < 0.01$). Notably, the strongest effects are observed in the 2–4 s range for MPIIGroupInteraction and in the 3–5 s range for CCDb, indicating that directional motion encoding

offers consistent and reliable benefits across varying temporal contexts. The consistent significance across datasets strengthens the conclusion that motion-differential preprocessing not only enhances model performance empirically but also provides statistically verifiable advantages in detecting subtle listener behaviors.

Impact of Attention Modules. The hierarchical attention architecture of TMAN is pivotal in capturing the multiscale dynamics of listener behavior. Each module contributes a distinct yet complementary analytical function. The *Micro-CM* module performs local cross-channel reasoning, identifying how fine-grained micro actions—such as simultaneous head nods and gaze shifts—co-occur across behavioral streams. The *UTM* module models temporal dependencies within each channel, capturing how motion and expression evolve across frames. The *Macro-CM* module subsequently integrates these temporally encoded signals, adaptively weighting each modality according to its contextual reliability. The ablation analyses confirm that each component individually improves baseline performance, while their integration produces a synergistic effect. This hierarchical structure enables TMAN to represent both short, reactive gestures and longer, continuous expressions within a unified modeling framework. From a cognitive perspective, this design mirrors the hierarchical structure of human perception, where localized sensory cues are first processed independently and later combined into higher-level perceptual judgments. Analogously, TMAN's bottom-up attention flow—from micro to macro—allows the model to distinguish between coincidental movement and socially meaningful responses. The consistent improvements observed when combining the three modules validate the necessity of such multilevel integration for decoding socially contingent, temporally distributed behaviors like backchanneling.

Impact of Visual Channels. The comparative evaluation of single and combined visual channels reveals that listener feedback is encoded through a distributed network of bodily, facial, and ocular signals, each contributing distinct yet complementary information. Among individual channels, head movement and rotation are the

most discriminative indicators of backchanneling, achieving accuracies of 74.4% on the MPIIGroupInteraction dataset and 79.4% on CCDb. Their dominance reflects the central role of nodding and re-orientation in signaling acknowledgment and understanding. Facial action units further refine this signal by conveying affective valence and the intensity of agreement—subtly differentiating neutral nods from affirming smiles. Body movement contributes low-frequency postural adjustments that ground these micro-actions within broader displays of attentiveness, while gaze direction encodes moment-to-moment shifts in attention and turn anticipation. Integrating all four channels yields the highest performance across both datasets, confirming that effective recognition of listener feedback requires the joint modeling of expressive, attentional, and postural components. The Macro-CM module's adaptive weighting behavior aligns with these findings: in dyadic CCDb interactions, attention concentrates on facial and head channels, whereas in the wider-angle MPIIGroupInteraction recordings, body and gaze cues gain relative importance. This synergy demonstrates that backchanneling is not localized to a single modality but emerges from the coordinated dynamics of multiple behavioral layers—an insight that underpins TMAN's cross-channel attention design.

Effectiveness and Internal Mechanism of Attention Modules in TMAN Inference. The qualitative case studies in Figures 3.7–3.8 shed light on how the three attention modules in TMAN work together to interpret multimodal behaviour. Alongside the quantitative results, they help explain why the model performs consistently across datasets and conditions. At the first stage, the *Micro-CM* module filters and strengthens the most relevant visual cues. When a listener nods, the model increases attention on the head and gaze channels while reducing the weight of unrelated motion. This early filtering enables the network to focus on small but meaningful movements that usually indicate attentiveness. The *UTM* module then determines which moments within the sequence are most informative. It assigns higher weights to frames corresponding to the start or peak of a backchannel, allowing the model to

capture both brief reactions and slower, continuous gestures. This mechanism helps TMAN distinguish between transient background motion and genuine listener feedback. Finally, the *Macro-CM* module integrates the information from all behavioural channels and adjusts their relative importance according to the visual context. When head cues are clear and dominant, Macro-CM relies mainly on them; when they are occluded or ambiguous, the model shifts attention to facial or gaze cues. This flexibility makes TMAN more robust to occlusion, camera angle, and inter-individual variation. Overall, these modules form a coherent inference process in which local features are refined, temporally organised, and adaptively combined. The correspondence between attention patterns and observable listener behaviours suggests that TMAN bases its decisions on meaningful social cues rather than on spurious correlations. This interpretability is valuable for understanding why the model succeeds or fails in specific cases and for increasing trust in its predictions. Compared with general-purpose vision-language models, TMAN's attention mechanisms are explicitly designed for social signal understanding. The model's ability to handle short, overlapping, and context-dependent cues explains its advantage in detecting subtle nonverbal feedback. These findings indicate that domain-specific attention designs remain necessary for tasks where the temporal and relational structure of behaviour carries the key to understanding.

3.8 Summary

This chapter presented TMAN, a Temporal Multichannel Attention Network designed for backchannel detection in group conversations. The model employs a hierarchical attention architecture consisting of the Micro-CM, UTM, and Macro-CM modules to capture both fine-grained and long-term behavioral patterns. Comprehensive experiments on the MPIIGroupInteraction and CCDb datasets demonstrated that TMAN achieves strong and consistent performance, surpassing existing state-of-the-art meth-

ods. The ablation results further confirmed that each attention module contributes distinct benefits, while motion-focused feature transformations—especially the temporal subtraction representation—significantly improve detection accuracy. The interpretability analyses provided additional insights into how TMAN allocates attention across time and modalities, explaining its robustness to diverse listener backchannel behaviors. Overall, this study emphasizes the importance of integrating temporal modeling and multimodal attention for understanding nonverbal communication in conversation. The next chapter extends this framework toward modeling higher-level social alignment by estimating agreement intensity from multimodal backchannel behaviors.

Chapter 4

Agreement Estimation from Backchannel Behaviours

In Chapter 3, we investigated the detection of backchannel responses in group conversation, focusing on identifying listeners' subtle nonverbal signals—such as head nods, gaze shifts, and facial expressions—that convey attentiveness and understanding. While recognizing the occurrence of backchannels is a fundamental step toward modelling interactive responsiveness, these cues can serve multiple communicative functions beyond maintaining conversational flow and turn-taking. Depending on context, a backchannel may signal acknowledgment [102], encouragement for the speaker to continue [42], or reflect a more affective and cognitive response such as agreement or disagreement with what has been said [35, 19]. Among these functions, the expression of agreement represents one of the most salient and socially meaningful aspects that backchannels can convey.

It is important to note that "agreement" as a communicative phenomenon encompasses a range of related but distinct acts. In the communication and pragmatics literature, scholars have distinguished among acknowledgment (recognising that a message has been received), alignment (adopting a similar stance or perspective),

endorsement (explicitly supporting a position), compliance (yielding to a request or expectation), and affective support (expressing empathy or solidarity) [88, 77]. These distinctions carry different implications for social dynamics: endorsement implies a stronger cognitive commitment than acknowledgment, while compliance may reflect social pressure rather than genuine concurrence. In the present study, the agreement annotation captures a composite perception of the listener's attitudinal alignment as judged by external annotators, without decomposing it into these finer-grained communicative functions. The continuous scale from -1 (total disagreement) to $+1$ (full agreement) therefore represents the annotators' holistic impression of the listener's stance, which may blend elements of cognitive alignment, affective warmth, and social courtesy. This operationalisation is consistent with prior computational work on agreement estimation [66, 19], but it is important to recognise that the resulting models capture perceived agreement intensity rather than a specific communicative act. Future work could benefit from more differentiated annotation schemes that separate these dimensions, enabling models to predict not only the degree but also the type of agreement being expressed.

This chapter therefore extends the analysis from identifying the presence of backchannel responses to estimating the degree of agreement they express, investigating how multimodal listener cues reveal varying levels of agreement during conversation. Specifically, we aim to quantify the extent to which visual feedback behaviours—such as nodding, smiling, or gaze coordination—reflect the listener's affective and cognitive alignment with the speaker. Building upon the temporal and cross-channel modelling framework established in Chapter 3, we introduce methods for estimating agreement intensity from multimodal backchannel behaviours. This task not only deepens our understanding of how social consensus is expressed nonverbally but also provides an important step toward developing intelligent systems capable of interpreting subtle interpersonal dynamics in human communication.

The remainder of this chapter is structured as follows. Section 4.1 introduces the

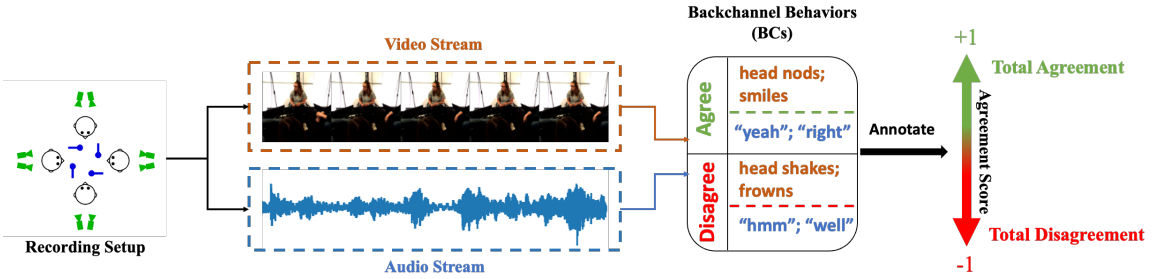
dataset and annotation protocol used for agreement estimation. Section 4.2 details the extraction of multimodal behavioural features from audio and video streams. Section 4.3 presents the proposed hierarchical framework for multimodal agreement estimation. Section 4.4 reports the experimental setup and evaluation results, while Section 4.5 conducts ablation studies examining the influence of different features, preprocessing strategies, and speaker diarization. Section 4.6 discusses the empirical findings and their broader implications. Finally, Section 4.7 summarises the key insights and outlines directions for subsequent research.

4.1 Dataset

This study employs the *MPIIGroupInteraction* dataset [67, 66], which provides synchronized audiovisual recordings of natural multi-party discussions and includes annotations for both the presence of backchannel responses and the degree of agreement they express. The dataset comprises 22 group discussions, each lasting approximately 20 minutes, amounting to over 440 minutes of interaction data. A total of 78 participants (43 female, 35 male), aged between 18 and 38, were recruited from a German university campus, all native German speakers to ensure linguistic consistency. Each group consisted of either three or four participants (10 groups of three and 12 groups of four), seated in a circular arrangement to allow mutual visibility.

As shown in Figure 4.1, the conversations were captured using inward-facing cameras and microphones, providing synchronized video and audio streams for each participant. Each recording was manually segmented into ten-second clips, where the final second corresponds to a listener's backchannel response. Annotators examined both the visual and auditory modalities to determine the extent of agreement or disagreement expressed in each response. Specifically, head nods, smiles, and affirmative vocalizations (e.g., "yeah," "right") were considered indicators of agreement, while head shakes, frowns, and hesitant utterances (e.g., "hmm," "well") were associated with

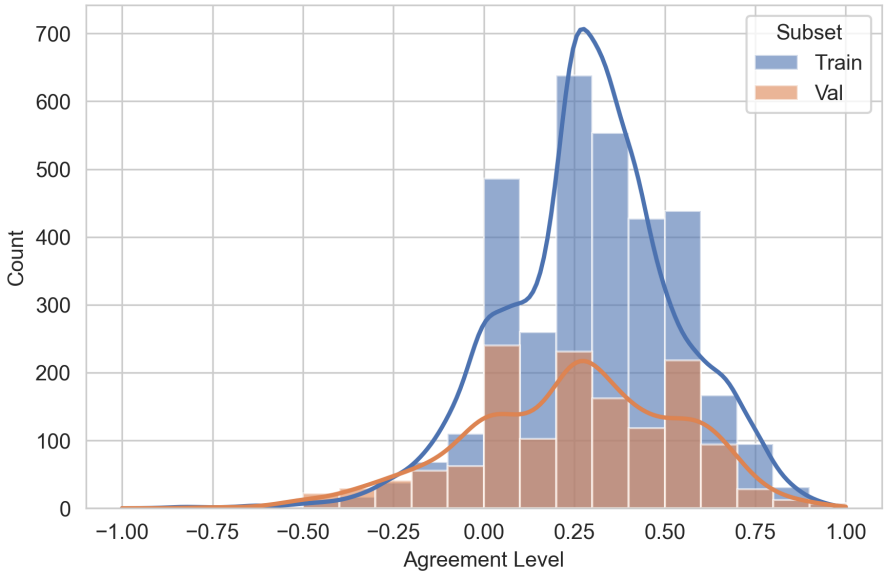
Figure 4.1 Illustration of the dataset and the annotation scheme: multimodal recordings of group conversations are captured using synchronized video and audio streams from multiple cameras and microphones. Annotators assign continuous agreement scores ($y \in [-1, 1]$) based on visual and auditory backchannel cues



disagreement. The perceived intensity of agreement was rated on a continuous scale $y \in [-1, 1]$, where $+1$ represents full agreement, -1 represents total disagreement, and 0 denotes a neutral or ambiguous reaction. This annotation scheme captures both binary and graded expressions of listener alignment, enabling more fine-grained modeling than categorical labels.

To ensure reliability, three trained annotators independently reviewed all candidate backchannel segments. If any annotator judged a segment as incorrectly labeled or ambiguous, it was flagged for further review. Segments unanimously deemed incorrect were removed from the dataset. Inter-annotator consistency yielded a Cohen's κ of 0.62 [66], indicating substantial agreement. The final ground-truth label for each sample was obtained by averaging the three annotators' ratings. This process resulted in a total of 7,234 validated backchannel instances, which were divided into 3,358 training samples and 1,427 validation samples. Each instance includes a ten-second video clip and its aligned audio waveform, with the target value representing the mean agreement intensity of the final second. The overall distribution of annotated agreement levels is shown in Figure 4.2. The majority of samples cluster in the mildly positive range (0–0.5), reflecting that agreement occurs more frequently than

Figure 4.2 Distribution of annotated agreement scores in the training and validation sets. Most samples fall within the mildly positive range (0–0.5), indicating that agreement responses are more frequent than disagreement in natural conversation.



disagreement in natural conversation. Instances of strong disagreement (below 0) are comparatively rare, which presents a notable challenge for model training and evaluation. This imbalance underscores the subtle nature of disagreement cues in spontaneous social interaction and highlights the need for robust multimodal modeling strategies capable of capturing both overt and nuanced listener feedback.

4.2 Feature Extraction

For agreement estimation, we extract multimodal behavioural features from both the visual and auditory modalities to capture the expressive cues underlying different degrees of agreement. Whereas Chapter 3 focused on detecting the occurrence of visual backchannel responses, this chapter aims to estimate their semantic orientation—whether a listener's reaction expresses agreement, neutrality, or disagreement.

Such interpretation often requires integrating information across modalities, as agreement is conveyed not only through visible movements but also through prosodic and vocal characteristics that reflect the speaker's affective stance. Therefore, the present task adopts a multimodal feature representation combining both video and audio streams. Furthermore, we employ the Whisper [80] to automatically transcribe the audio into text, from which additional linguistic features are extracted for analysis.

Speaker Diarization

Before multimodal feature extraction, it is essential to accurately identify when each participant is speaking, as this allows us to distinguish the true source of vocal activity in multi-party recordings. In the recording setup, multiple tabletop microphones are used to capture the group discussion, resulting in mixed audio signals that contain overlapping voices from all participants without explicit channel separation. Consequently, without reliable speaker identification, it becomes difficult to determine whether a particular sound or utterance originates from the target individual or another participant. This distinction is crucial, as brief vocalizations from listeners—such as short utterances or paraverbal cues—often form part of backchannel behaviour and convey varying degrees of agreement or engagement. Accurate detection of speaking moments therefore ensures that the extracted audio-visual and linguistic features correspond to the correct speaker, preventing cross-contamination between speaker and listener behaviours. To achieve this, we employ an audio-visual active speaker detection (ASD) model [56] that robustly identifies frame-level speaking activity for each participant.

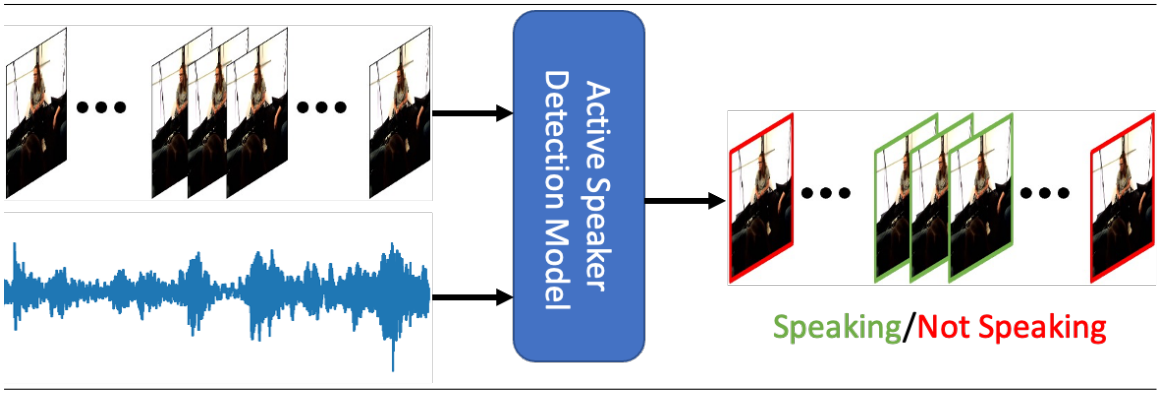
The ASD model integrates facial motion and acoustic energy to determine speaking status at the frame level. Given a sequence of video frames $\mathcal{V} = \{v_1, v_2, \ldots, v_T\}$ and the corresponding audio stream $\mathcal{A} = \{a_1, a_2, \ldots, a_T\}$, the ASD network f_{ASD} predicts

a binary label for each frame:

$$s_t = f_{\text{ASD}}(v_t, a_t), \quad s_t \in \{0, 1\}, \tag{4.1}$$

where $s_t = 1$ denotes that the target individual is speaking at time t, and $s_t = 0$ indicates silence or listening behaviour. The model jointly analyses visual cues such as lip motion, facial dynamics, and temporal consistency, alongside audio features such as short-term energy and speech activity, to infer s_t robustly even under conditions of overlapping speech or background noise.

Figure 4.3 Illustration of speaking instances identification using audio-visual signals and active speaker detection. Each frame is classified as "Speaking" (green) or "Not Speaking" (red) for the target individual.



As illustrated in Figure 4.3, each frame is classified as either "Speaking" (green) or "Not Speaking" (red). The resulting binary sequence forms a temporal mask that delineates speaker activity throughout the conversation. Only the frames labelled as speaking ($s_t = 1$) are retained for the extraction of audio-visual features associated with the speaker's utterances, whereas the remaining frames are reserved for analysing listener behaviour and potential backchannel responses. This diarization process ensures that all subsequent multimodal representations are temporally aligned with the correct speaker identity, thereby enhancing the precision of agreement estimation in multi-party conversation.

Audio Features

Acoustic signals provide rich paralinguistic information that complements visual behaviour in agreement estimation. While visual cues reflect overt gestures and facial expressions, audio features capture subtle prosodic variations—such as pitch, loudness, and rhythm—that convey the listener's affective state and level of endorsement. For instance, an affirmative "yeah" is often produced with higher pitch and smoother intonation than a hesitant "hmm" or a reluctant "well." These vocal modulations reveal the emotional and attitudinal dimension of agreement that may not be observable through facial or bodily motion alone.

Low-level Acoustic Features. To quantify these paralinguistic cues, we extract features using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [32, 31]. eGeMAPS provides a compact yet perceptually meaningful representation of vocal expression, resulting in a d_e-dimensional feature vector per analysis window. Each feature group captures a distinct aspect of the speech signal relevant to affective and interactional analysis:

- **Frequency-related features:** Fundamental frequency (F_0), jitter, and formant frequencies (F1–F3) capture pitch contour and vocal stability. Higher average F_0 and reduced jitter are typically associated with positive affect and agreement, whereas increased irregularity or pitch lowering may indicate hesitation or disagreement.

- **Energy-related features:** Loudness, root-mean-square (RMS) energy, and shimmer describe amplitude dynamics and vocal effort. Listeners expressing agreement often produce more energetic and continuous vocal responses, while lower intensity or abrupt amplitude changes may correspond to weak endorsement or disengagement.

- **Spectral-related features:** Spectral flux, centroid, and slope reflect timbral and articulatory properties that contribute to the perceived tone and warmth of the voice. These parameters help distinguish soft, supportive backchannels (e.g., "mm-hmm") from sharper or more abrupt responses conveying disapproval or contrast.

All acoustic features are extracted using the openSMILE toolkit [32] with a one-second window and a stride corresponding to 30 frames per second, ensuring that one audio feature vector is temporally aligned with each video frame. This synchronization ensures that the temporal evolution of vocal and visual signals can be jointly modelled within the multimodal framework.

Pretrained Audio Embeddings. In addition to handcrafted acoustic descriptors, we extract high-level audio representations using the pretrained WavLM model [22]. WavLM is a transformer-based self-supervised speech model trained on large-scale speech corpora to capture both phonetic and prosodic information across diverse acoustic environments. Unlike eGeMAPS, which encodes predefined low-level signal properties, WavLM embeddings represent abstract latent features that integrate contextual, temporal, and affective characteristics of speech. These embeddings provide a richer description of vocal tone and delivery, enabling the model to differentiate subtle variations in listener responses such as confident agreement, uncertain acknowledgment, or mild disagreement. For each audio segment, WavLM processes the waveform and outputs frame-level embeddings at 50 Hz, which are subsequently downsampled to 30 frames per second to maintain alignment with visual features. These pretrained representations complement the low-level eGeMAPS features by capturing higher-order patterns of intonation and prosody, thereby enhancing the model's ability to infer the affective and semantic orientation of backchannel utterances.

Video Features

Visual behaviours play a crucial role in signalling agreement during conversation, as subtle facial and bodily cues often reveal a listener's evaluative stance even in the absence of speech. While Chapter 3 focused on visual backchannel detection to identify the presence of listener feedback, the present task extends this analysis to quantify the degree of agreement conveyed by such behaviours. In this study, we extract multiple video-based features to capture both macro-level motion and micro-level facial expressions that correspond to different levels of attitudinal alignment.

Facial Behaviour Features. In this study, we employ OpenFace 2.0 [11] to extract a comprehensive set of facial features, including eye gaze, head orientation, facial landmarks, and facial action units (AUs):

- ***Eye gaze:*** OpenFace estimates the three-dimensional gaze direction vector (x, y, z) for each eye and provides averaged horizontal and vertical components in radians. Gaze orientation reflects the listener's visual focus and attentional engagement with the speaker. Sustained eye contact or synchronized gaze shifts often indicate agreement or active listening, while gaze aversion may signal hesitation or mild disagreement.

- ***Head pose:*** Six parameters are extracted for head position and rotation in three-dimensional space: translational coordinates (x, y, z) in millimetres and rotational angles (pitch, yaw, roll) in degrees. Head nods, tilts, and shakes constitute primary nonverbal indicators of agreement and disagreement. Positive agreement is frequently expressed through rhythmic nodding or forward inclination, whereas lateral or downward head movements may convey reservation or disapproval.

- ***Facial action units (AUs):*** Seventeen facial AUs are extracted according to the Facial Action Coding System (FACS) [30], each representing the activation

intensity of specific muscle groups on a continuous scale from 0 (inactive) to 5 (strongly active). Relevant indicators of agreement include AU6 (cheek raiser) and AU12 (lip corner puller), which correspond to smiling, while AUs such as AU15 (lip corner depressor) or AU17 (chin raiser) may accompany disagreement or uncertainty.

- ***Facial landmarks:*** OpenFace detects 68 two-dimensional facial landmarks corresponding to key points around the eyes, eyebrows, nose, mouth, and jawline. These coordinates describe the geometric configuration of the face and capture subtle spatial deformations associated with expressions of agreement or disagreement. For example, upward movement of the lip corners, eyebrow raises, or changes in mouth shape can indicate positive alignment, while tightened lips or downward mouth curvature may signal disagreement or uncertainty. The landmark coordinates are normalised with respect to the bounding box size and centred at the nose tip to ensure scale and translation invariance across frames and participants.

Body Pose Features. In addition to facial and gaze cues, upper-body movement provides important contextual information for interpreting agreement and disagreement. Body motion often complements facial expressions, reinforcing or modulating the listener's affective stance. Following the procedure in Chapter 3, we employ ViTPose [101], a vision transformer–based pose estimator, to extract two-dimensional skeletal keypoints from each video frame. ViTPose detects 17 body joints in the COCO format [57], from which we retain 11 upper-body keypoints—nose, eyes, ears, shoulders, elbows, and wrists—since these regions are consistently visible and most relevant to backchannel and agreement behaviours in seated conversations. The spatial coordinates (x, y) of these joints are normalised with respect to the bounding box size to ensure scale invariance across participants and camera views. The resulting joint trajectories capture both global and local body movements that accompany

attitudinal expressions.

Frame Embedding Features. To complement low-level behavioural descriptors, we extract high-level semantic representations from raw video frames using a pretrained Vision Transformer (ViT) model [28]. While features from OpenFace and ViTPose capture explicit motion and expression dynamics, ViT embeddings provide a holistic encoding of visual appearance and spatial configuration, enabling the model to capture contextual and compositional cues that may underlie agreement-related behaviours. The inclusion of such high-level visual embeddings allows the framework to integrate both interpretable behavioural features and abstract visual semantics within a unified representation. Specifically, we employ the ViT-B/16 architecture pretrained on ImageNet [27] to generate frame-level embeddings. Each frame is resized to 224×224 pixels and divided into non-overlapping patches, which are linearly projected and processed through transformer layers to yield a d_v-dimensional feature vector. These embeddings capture spatial relations among facial regions, body posture, and background context without relying on explicit landmark detection.

The extracted features across all modalities are summarised in Table 4.1. Each modality provides complementary information reflecting different behavioural dimensions of agreement. Audio features capture prosodic and spectral variations that reveal vocal tone and emphasis, while video features encode facial expressions, gaze dynamics, and postural movement indicative of social alignment. High-level embeddings from WavLM and ViT further enrich these handcrafted descriptors by providing abstract, context-aware representations. Together, these multimodal features form the input representation for subsequent temporal modelling and agreement estimation.

Table 4.1: Feature categories, feature names, and their dimensions extracted from each modality.

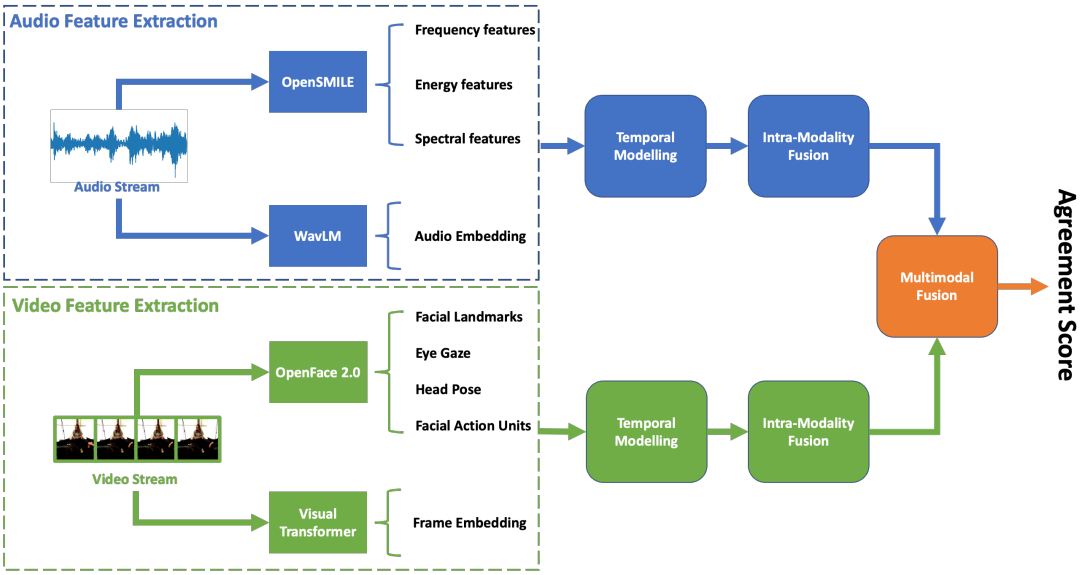
Modality	Category	Features	Dimension
Audio	Acoustic Features	Frequency-related	20
		Energy-related	21
		Spectral-related	47
	Speech Embedding	WavLM Embedding	768
Video	Facial Behaviors	Facial Landmarks	136
		Eye Gaze	6
		Head Pose	6
		Facial Action Units	35
	Body Pose	Body Keypoints	22
	Frame Embedding	Pretrained ViT Embedding	768

4.3 Methodology

This study proposes a hierarchical multimodal framework for estimating the degree of agreement expressed during group conversations. The architecture is designed to model both the temporal evolution and the cross-modal interplay of behavioural and acoustic cues that signal agreement or disagreement. As illustrated in Figure 4.4, the model processes audio and video streams through three successive stages: (1) **Temporal Modelling**, which encodes the short-term dynamics within each modality; (2) **Intra-Modality Fusion**, which integrates heterogeneous features within the same modality to produce unified modality-level representations; and (3) **Multimodal Fusion**, which captures interdependencies between modalities to estimate a continuous agreement score. Through this hierarchical design, the system learns to combine com-

plementary cues such as prosody, facial expression, and body motion in an adaptive and context-aware manner.

Figure 4.4 Overview of the proposed multimodal agreement estimation framework. Audio and video features are first processed independently through temporal modelling and intra-modality fusion. The resulting modality-specific embeddings are then integrated through multimodal attention to predict the final agreement intensity.



Temporal Modelling. The first stage of the framework focuses on learning the temporal dependencies that characterise dynamic behavioural patterns within each modality. Let $\mathcal{X}^{(m)} = \{x_1^{(m)}, x_2^{(m)}, \ldots, x_T^{(m)}\}$ denote the feature sequence of modality $m \in \{\text{audio}, \text{video}\}$, where T represents the number of frames or time steps in a ten-second segment. Each sequence is passed through a Gated Recurrent Unit (GRU) network, which captures both short- and long-range temporal correlations by

recursively updating its hidden state:

$$r_t^{(m)} = \sigma(W_r^{(m)} x_t^{(m)} + U_r^{(m)} h_{t-1}^{(m)}), \tag{4.2}$$

$$z_t^{(m)} = \sigma(W_z^{(m)} x_t^{(m)} + U_z^{(m)} h_{t-1}^{(m)}), \tag{4.3}$$

$$\tilde{h}_t^{(m)} = \tanh(W_h^{(m)} x_t^{(m)} + r_t^{(m)} \odot U_h^{(m)} h_{t-1}^{(m)}), \tag{4.4}$$

$$h_t^{(m)} = (1 - z_t^{(m)}) \odot h_{t-1}^{(m)} + z_t^{(m)} \odot \tilde{h}_t^{(m)}, \tag{4.5}$$

where $x_t^{(m)}$ and $h_t^{(m)}$ denote the input and hidden state of modality m at time step t, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The GRU transforms the input sequence into a temporally encoded representation $H^{(m)} = \{h_1^{(m)}, \ldots, h_T^{(m)}\}$ that preserves contextual information over time.

To emphasise the most informative temporal regions, we apply a temporal attention mechanism that weights each hidden state according to its relevance:

$$\alpha_t^{(m)} = \text{softmax}(w_a^{(m)\top} \tanh(W_a^{(m)} h_t^{(m)} + b_a^{(m)})), \tag{4.6}$$

$$c^{(m)} = \sum_{t=1}^{T} \alpha_t^{(m)} h_t^{(m)}, \tag{4.7}$$

where $\alpha_t^{(m)}$ represents the attention weight at time step t, and $c^{(m)} \in \mathbb{R}^h$ is the context vector summarising the temporal evolution of modality m. This mechanism enables the network to focus on key moments, such as short affirmative vocalisations or expressive gestures, that are most indicative of agreement.

Intra-Modality Fusion. Each modality comprises multiple feature sources that provide complementary information. For instance, the audio modality includes low-level acoustic descriptors (eGeMAPS) and high-level speech embeddings (WavLM), while the video modality combines interpretable behavioural cues (OpenFace, ViT-Pose) with high-level semantic embeddings (ViT). To integrate these within-modality signals, we employ an attention-based fusion mechanism that learns to balance their relative importance dynamically.

Given a set of K_m feature embeddings for modality m, $\{c^{(m,1)}, c^{(m,2)}, \ldots, c^{(m,K_m)}\}$, we compute their weighted combination as:

$$\beta_k^{(m)} = \text{softmax}(w_f^{(m)\top} \tanh(W_f^{(m)} c^{(m,k)} + b_f^{(m)})), \tag{4.8}$$

$$f^{(m)} = \sum_{k=1}^{K_m} \beta_k^{(m)} c^{(m,k)}, \tag{4.9}$$

where $\beta_k^{(m)}$ denotes the attention weight of the k-th feature source, and $f^{(m)}$ is the fused intra-modality representation. This mechanism allows the model to adaptively emphasise the most informative feature types within each modality—for example, prioritising prosodic patterns when vocal tone dominates agreement cues, or giving more weight to facial and postural movements when visual expressiveness is stronger.

Multimodal Fusion. The final stage integrates the modality-specific embeddings $\{f^{(\text{audio})}, f^{(\text{video})}\}$ to capture cross-modal interactions that jointly convey agreement. Rather than treating modalities independently, we use a multimodal attention mechanism that aligns and fuses their latent representations:

$$\gamma_m = \text{softmax}(w_M^\top \tanh(W_M f^{(m)} + b_M)), \tag{4.10}$$

$$f^* = \sum_m \gamma_m f^{(m)}, \tag{4.11}$$

where γ_m denotes the modality-level attention weight and f^* represents the fused multimodal embedding. This step enables the model to learn synergistic relationships, such as how a smile reinforces a positive vocal tone, or how mismatched cues—like neutral facial expression with a hesitant tone—reduce the perceived agreement.

Finally, the fused representation is passed through a regression head composed of two fully connected layers with nonlinear activation to predict a continuous agreement score:

$$\hat{y} = W_2 \,\text{ReLU}(W_1 f^* + b_1) + b_2, \tag{4.12}$$

where $\hat{y} \in [-1, 1]$ denotes the estimated agreement intensity, with higher values indicating stronger agreement. The network is trained using the mean squared error

(MSE) loss between the predicted and annotated agreement scores:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \tag{4.13}$$

where N is the number of training samples and y_i is the ground-truth label.

4.4 Results

This section presents the experimental evaluation of the proposed multimodal framework for agreement estimation. We first describe the experimental setup and training configuration, followed by an analysis of the optimal model settings. Comprehensive ablation studies are then conducted to examine the influence of different backbone architectures, observation window lengths, and speaker diarization strategies, as well as the contribution of individual modalities and the effectiveness of the proposed multimodal integration scheme.

4.4.1 Experimental Setup

Feature extraction follows the procedures described in Section 4.2, using both audio and visual modalities. Acoustic features are derived from eGeMAPS and WavLM to capture low-level prosodic attributes and high-level contextual representations, while visual features are obtained from OpenFace 2.0, ViTPose, and the Vision Transformer (ViT) to represent facial, postural, and spatial–semantic cues. All features are temporally aligned at 25 frames per second to ensure consistent synchronisation across modalities. All models are implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU with 24 GB of memory. Each model is trained for up to 150 epochs with a batch size of 64. The Adam optimiser is used with an initial learning rate of 0.0005 and a weight decay of 0.0001. Within the temporal modelling component, a single-layer GRU with 64 hidden units and a dropout rate of 0.3 is

employed. Layer normalisation is applied after the GRU, and additional dropout layers (with a rate of 0.5) are inserted after fusion modules to mitigate overfitting. The hidden dimension h is set to 128 for temporal encoding and 64 for subsequent integration layers. The network is optimised using the mean squared error (MSE) loss between predicted and ground-truth agreement scores. To ensure fair comparison, all experiments are conducted on the training and validation splits of the MPIIGroupInteraction dataset, following the continuous agreement annotation protocol described in Section 4.1. Model performance is evaluated using mean squared error (MSE) as the sole metric to assess the precision of regression-based agreement estimation.

4.4.2 Evaluation of the overall framework

We evaluate the proposed multimodal framework for agreement estimation against several established baselines and recent state-of-the-art approaches. The comparison covers both unimodal and multimodal configurations using the MPIIGroupInteraction dataset, with results reported in terms of the best validation mean squared error (MSE). As shown in Table 4.2, our framework consistently outperforms previous methods across all evaluated modalities.

The baseline models adapted from Müller et al. [66] represent traditional unimodal and multimodal fusion schemes based on handcrafted descriptors such as head pose and eGeMAPS. These approaches show limited ability to capture the temporal and contextual subtleties of agreement behaviour, yielding validation MSEs between 0.075 and 0.085. Sharma et al. [83] employ graph-based fusion of deep audio-visual features, while Amer et al. [2] use OpenFace and OpenPose for visual cue extraction; both obtain moderate improvements, with the latter achieving an MSE of 0.064. The proposed framework yields further improvements, particularly in the multimodal setting. The video-based model combining facial landmarks, head pose, and facial action units attains an MSE of 0.054, outperforming all previous visual methods in-

Table 4.2: Comparison of different methods, modalities, and feature configurations for agreement estimation on the MPIIGroupInteraction dataset. Results are reported in terms of the best validation mean squared error (MSE).

Method	**Modality**	**Features**	**Best Val. MSE**
Müller et al. [66]	Video	Head Pose	0.075
	Audio	eGeMAPS	0.085
	Audio+Video	All features	0.079
Sharma et al. [83]	Audio+Video	Deep features	0.075
Amer et al. [2]	Video	OpenFace + OpenPose	0.064
Ours	Video	Facial Landmarks+Head Pose+Facial AUs	0.054
	Audio	Energy+Spectral	0.076
	Audio+Video	Facial Landmarks+Head Pose+Facial AUs +Energy+Spectral	**0.052**

cluding Amer et al. (0.064). Integrating audio and video modalities further improves performance, with the configuration combining visual features and selected acoustic descriptors (energy and spectral features) attaining the lowest validation MSE of 0.052. These results suggest that acoustic and visual cues provide complementary information for estimating listener agreement. The contribution of speaker diarization to both modalities is examined in detail in Section 4.5. Overall, the findings indicate that the proposed framework benefits from the joint modelling of prosodic and behavioural cues, and that incorporating speaker information enhances the reliability of multimodal agreement estimation. The integration of temporal encoding with adaptive multimodal fusion appears to facilitate a more consistent estimation of subtle agreement variations in natural conversational settings.

4.5 Ablation Studies

To understand the contribution of each modality and modeling component to the prediction of agreement intensity, this section presents a series of ablation experiments conducted on the proposed framework. These analyses aim to disentangle how distinct feature types, temporal modeling strategies, preprocessing methods, and role segmentation mechanisms influence model performance. By systematically isolating and evaluating each factor, we provide quantitative evidence for the design choices made in the final multimodal architecture. The ablation is organized into three complementary parts. First, we investigate the **audio modality**, examining how different acoustic descriptors, backbone architectures, and temporal preprocessing strategies affect estimation accuracy. The analysis reveals that prosodic energy and spectral cues offer the most reliable audio signals for agreement modeling. Second, we focus on the **video modality**, evaluating various facial, head, and body representations. By applying temporal differencing and selective intra-modality fusion, this analysis demonstrates that fine-grained facial dynamics substantially outperform holistic frame embeddings and body posture features. Finally, we assess the impact of **speaker diarization**, comparing results obtained with and without explicit separation of speaker roles. This experiment clarifies how isolating listener-specific segments improves both audio and visual modalities by reducing interference from non-listener activity. These ablation studies establish the relative importance of different modalities and processing strategies. The results confirm that short-term visual dynamics and energy-based acoustic variations are the dominant indicators of listener agreement, while speaker diarization provides an additional refinement that enhances signal clarity and model interpretability.

4.5.1 Audio Features

We examine the performance of the audio modality through systematic analyses of unimodal feature types, backbone architectures, feature engineering methods, tem-

Table 4.3: Evaluation of audio features

Category	Feature	Model	Feature Engineering	1s	2s	3s	4s	5s
Acoustic Features	Energy	GRU	Original	0.084	0.083	0.082	0.081	**0.079**
			Subtraction	0.084	0.082	0.082	0.080	**0.081**
		Transformer	Original	0.086	0.085	0.084	**0.081**	0.082
			Subtraction	0.085	0.086	0.086	**0.083**	**0.083**
	Frequency	GRU	Original	0.085	0.084	0.083	0.083	**0.082**
			Subtraction	0.085	0.083	**0.082**	0.083	0.083
		Transformer	Original	0.085	0.084	0.084	0.083	**0.082**
			Subtraction	0.086	0.084	0.084	0.083	**0.083**
	Spectral	GRU	Original	0.084	0.082	0.083	0.082	**0.080**
			Subtraction	0.083	0.083	0.083	0.083	**0.082**
		Transformer	Original	0.085	0.084	0.083	0.083	**0.082**
			Subtraction	0.088	0.086	0.086	0.084	**0.083**
Speech Embedding	WavLM Embedding	GRU	Original	0.084	0.084	0.083	0.081	**0.080**
			Subtraction	0.083	0.084	0.082	0.082	**0.081**
		Transformer	Original	0.083	0.084	0.082	0.082	**0.081**
			Subtraction	0.084	0.084	0.083	0.083	**0.082**

poral window lengths, and intra-modality fusion strategies. All results are evaluated using the validation mean squared error (MSE).

Backbone Architectures. Two sequence modeling architectures are compared: a gated recurrent unit (GRU) network [23] and a Transformer encoder [93]. The GRU models short-term temporal dependencies through recurrent gating, making it well suited for capturing smooth prosodic contours and local variations in energy and pitch. The Transformer, in contrast, employs self-attention to integrate information across longer temporal spans, allowing it to capture broader contextual patterns. As shown in Table 4.3, both architectures achieve comparable performance across feature

categories, yet the GRU consistently attains slightly lower MSE values—typically by about 0.002 to 0.004—across energy-, frequency-, spectral-, and embedding-based features. This steady advantage implies that agreement-related cues such as subtle modulations in loudness, pitch, and spectral composition are primarily short-range phenomena that can be effectively modeled through recurrent dynamics without requiring extended attention over distant frames. The Transformer performs competitively on high-level learned embeddings such as WavLM, where temporally aggregated representations benefit from attention-based integration, but it offers no clear advantage for handcrafted acoustic features.

Observation Windows. To determine the effective temporal span of relevant acoustic information, we vary the observation window from one to five seconds preceding each backchannel response. The results show a consistent decline in MSE with longer windows, followed by a plateau at around five seconds. For instance, the GRU with energy features improves from 0.084 at a one-second window to 0.079 at five seconds. Similar trends appear for frequency, spectral, and embedding features, all stabilizing around 0.080–0.082. These results suggest that the acoustic precursors of perceived agreement are short-term yet contextually accumulated within a few seconds before the listener's response. Extending the window beyond five seconds introduces minimal benefit, likely due to the inclusion of acoustically irrelevant information from preceding speech segments.

Feature Engineering. We further compare raw frame-level features with a *subtraction* transformation, which computes the temporal difference between consecutive frames to emphasize local changes in acoustic dynamics. However, as shown in Table 4.3, subtraction does not improve performance in this task. Across all feature categories, models trained on the original features consistently achieve slightly lower MSEs—by roughly 0.001–0.002—than their subtraction-based counterparts. This

pattern suggests that absolute acoustic magnitudes, rather than their temporal derivatives, provide more stable cues for estimating agreement intensity in audio signals. Unlike in backchannel detection (see Chapter 3), where rapid changes in energy or pitch serve as strong triggers, agreement-related cues appear to rely more on steady prosodic patterns than on abrupt variations.

Unimodal Feature Analysis. Among the four categories of audio features, energy-related descriptors achieve the best overall performance, with the GRU model attaining the lowest MSE of 0.079 at a five-second observation window. Spectral and WavLM features follow closely with MSEs of 0.080, while frequency-based features perform slightly worse at around 0.082. The narrow range of differences indicates that multiple aspects of the acoustic signal—energy dynamics, spectral composition, and high-level speech embeddings—each capture partially overlapping yet complementary cues of listener agreement. Energy-based features likely reflect the intensity and prosodic rhythm of speech that accompany agreement responses, whereas spectral descriptors encode timbral and resonance variations associated with emphasis or vocal tone. Meanwhile, WavLM embeddings encapsulate both prosodic and phonetic information through learned high-level representations. Overall, the audio modality alone yields moderate but consistent predictive power for agreement estimation, with both handcrafted and learned features contributing distinct, complementary perspectives on listener behavior.

Intra-Modality Fusion. We further examine whether combining multiple acoustic feature types within the audio modality enhances model performance. Using the optimal configurations from the unimodal experiments, several feature combinations are evaluated. As shown in Table 4.4, the fusion of *energy* and *spectral* features achieves the lowest validation MSE of 0.076, outperforming all individual features by approximately 0.003. This combination leverages the complementary strengths of both descriptors: energy features capture temporal variations in vocal intensity, whereas

Table 4.4: Evaluation of intra-modality fusion on the audio modality

Energy	**Frequency**	**Spectral**	**WavLM Embedding**	**Val. MSE**
✓	✓			0.079
✓		✓		**0.076**
✓			✓	0.080
✓	✓	✓		**0.078**
✓	✓		✓	0.080
	✓	✓	✓	0.081
✓	✓	✓	✓	**0.080**

spectral features encode frequency distribution and timbral texture. In contrast, adding additional features—such as frequency-based descriptors or WavLM embeddings—does not further reduce the error and occasionally degrades performance (MSE ≈ 0.080). The degradation likely arises from redundancy among high-dimensional representations and the limited contribution of less informative cues. These results indicate that targeted fusion of acoustically complementary features yields greater benefit than combining all available descriptors indiscriminately, emphasizing the importance of selective integration within a single modality.

4.5.2 Video Features

In addition to the audio modality, we evaluate the contribution of visual cues to agreement estimation. The same analytical structure is applied here—examining unimodal feature performance, temporal differencing, and intra-modality fusion—allowing direct comparison with the audio analysis in Section 4.5.1. Whereas acoustic features primarily encode prosodic and spectral variations, visual signals provide complementary information through observable facial and bodily dynamics that accompany subtle forms of listener feedback.

Table 4.5: Evaluation of video features.

Category	Feature	Model	Model Input	1s	2s	3s	4s	5s
Facial Behaviors	Facial Landmarks	GRU	Ori. Value	0.084	**0.083**	0.084	**0.083**	0.084
			Subtraction	0.063	0.059	**0.058**	**0.058**	0.059
		Transformer	Ori. Value	0.085	0.085	0.085	**0.083**	0.084
			Subtraction	0.068	0.061	**0.060**	0.064	0.065
	Eye Gaze	GRU	Ori. Value	**0.084**	0.085	**0.084**	**0.084**	**0.084**
			Subtraction	0.080	**0.078**	0.080	0.079	0.080
		Transformer	Ori. Value	0.085	0.085	0.085	**0.083**	0.085
			Subtraction	0.084	**0.080**	**0.080**	0.082	0.082
	Head Pose	GRU	Ori. Value	0.086	0.086	**0.085**	0.086	0.086
			Subtraction	0.075	0.076	**0.074**	0.077	0.077
		Transformer	Ori. Value	**0.085**	**0.085**	**0.085**	**0.085**	0.086
			Subtraction	0.081	0.079	**0.077**	0.078	0.079
	Facial Action Units	GRU	Ori. Value	**0.082**	0.083	0.083	0.083	0.083
			Subtraction	0.084	0.081	**0.080**	**0.080**	**0.080**
		Transformer	Ori. Value	**0.084**	**0.084**	0.085	0.086	**0.084**
			Subtraction	0.083	0.084	**0.082**	0.084	0.083
Body Pose	Body Keypoints	GRU	Ori. Value	**0.084**	**0.084**	**0.084**	**0.084**	**0.084**
			Subtraction	0.086	0.083	0.082	**0.081**	0.083
		Transformer	Ori. Value	**0.084**	**0.084**	**0.084**	**0.084**	0.085
			Subtraction	**0.084**	0.085	0.086	0.086	0.086
Frame Embedding	Pretrained ViT Embedding	GRU	Ori. Value	0.086	0.086	0.088	0.088	**0.085**
			Subtraction	**0.084**	**0.084**	**0.084**	**0.084**	0.085
		Transformer	Ori. Value	0.085	0.085	0.085	**0.084**	0.085
			Subtraction	0.086	**0.085**	**0.085**	0.086	0.087

Table 4.6: Evaluation of intra-modality fusion on the video modality

Facial Landmarks	Eye Gaze	Head Pose	Facial AUs	Body Keypoints	ViT Embedding	Val. MSE
✓	✓					0.060
✓		✓				**0.058**
✓			✓			**0.058**
✓				✓		0.068
✓					✓	0.078
	✓	✓				0.077
	✓		✓			0.071
	✓			✓		0.078
	✓				✓	0.081
		✓	✓			0.063
		✓		✓		0.077
		✓			✓	0.080
			✓	✓		0.078
			✓		✓	0.071
				✓	✓	0.081
✓	✓	✓				0.058
✓	✓		✓			**0.055**
✓	✓			✓		0.058
✓	✓				✓	0.059
	✓	✓	✓			0.069
	✓	✓			✓	0.071
	✓		✓		✓	0.073
		✓	✓	✓		0.067
		✓	✓		✓	0.077
			✓	✓	✓	0.078
✓	✓	✓	✓			**0.058**
✓	✓	✓		✓		0.062
✓	✓	✓			✓	0.078
	✓	✓	✓	✓		0.075
	✓	✓	✓		✓	0.079
		✓	✓	✓	✓	0.080
✓	✓	✓	✓	✓	✓	**0.081**

Backbone Architectures. Consistent with the audio experiments, both a gated recurrent unit (GRU) network and a Transformer encoder are evaluated for temporal modeling. The GRU, designed to capture local temporal dependencies, consistently yields lower MSEs than the Transformer across nearly all visual features, mirroring the pattern observed in the audio modality. This similarity reinforces that short-term temporal continuity—whether in vocal or visual behavior—plays a dominant role in conveying agreement cues, while long-range attention contributes little additional benefit.

Feature Engineering. We again examine the effect of framewise *subtraction*, which captures temporal differences between consecutive frames. The impact is markedly stronger in the visual modality than in the acoustic one. As shown in Table 4.5, subtraction substantially reduces prediction error for all facial behaviour features. The MSE for Facial Landmarks decreases from 0.083 to 0.058, and for Head Pose from 0.085 to 0.074. These improvements indicate that dynamic visual changes—micro-movements, nods, or brief gaze shifts—carry more discriminative information than static appearances. In contrast to the audio modality, where subtraction yielded negligible benefit, temporal differencing proves essential for highlighting the motion-based cues that accompany nonverbal agreement.

Unimodal Feature Analysis. Among the six visual descriptors, Facial Landmarks achieve the lowest validation MSE (0.058), outperforming all other video-based features and any unimodal audio counterpart (best audio MSE = 0.079). Head Pose and Eye Gaze follow with MSEs of approximately 0.074, reflecting their shared sensitivity to attentional orientation. Facial Action Units (AUs) yield 0.080, suggesting that local muscular activations, such as smiles or eyebrow raises, contribute moderately to agreement estimation. Body Keypoints (0.081) and ViT embeddings (0.084) perform least effectively, as coarse body posture and holistic frame appearance encode less

reliable affective information. Together, these results indicate that fine-grained facial dynamics are considerably more informative than global visual context, and more predictive than acoustic fluctuations alone.

Intra-Modality Fusion. To explore complementarity among visual cues, we conduct feature-level fusion within the video modality using the same strategy applied in audio fusion (Section 4.5.1). As reported in Table 4.6, the combination of *Facial Landmarks*, *Head Pose*, and *Facial AUs* achieves the lowest validation MSE of 0.055, improving upon the best unimodal result by about 0.003. This gain parallels the improvement seen in the optimal audio fusion (energy + spectral features, MSE = 0.076), though the visual fusion attains substantially lower overall error. The performance advantage arises from integrating geometric (landmarks and head orientation) and expressive (AUs) cues, which together capture both structural alignment and affective feedback. Adding additional streams—such as Eye Gaze, Body Keypoints, or ViT embeddings—does not enhance prediction accuracy and often leads to degradation (MSE $\approx$ 0.07–0.08), similar to the redundancy effect observed in the audio modality.

4.5.3 Speaker Diarization

To further examine how separating speaker roles affects modeling accuracy, we compare the performance of all acoustic and visual features with and without speaker diarization. This analysis aims to determine whether isolating listener-specific segments enhances the estimation of agreement intensity by reducing interference from irrelevant speech or visual cues belonging to other participants.

Audio Modality. Table 4.7 presents the comparison for audio features. Speaker diarization slightly improves performance across most handcrafted acoustic descrip-

Table 4.7: Evaluation of audio features with and without speaker diarization

Category	Feature	W/O Speaker Diarization	Speaker Diarization
Acoustic Features	Energy	0.079	**0.076**
	Frequency	0.082	**0.081**
	Spectral	0.080	**0.078**
Speech Embedding	WavLM Embedding	**0.080**	0.081

tors, with average reductions of 0.002–0.003 in validation MSE. The most notable gain is observed for the energy-based features (from 0.079 to 0.076), followed by spectral features (from 0.080 to 0.078). These improvements indicate that distinguishing who is speaking clarifies short-term prosodic fluctuations and removes misleading background speech segments that otherwise distort the estimation of listener agreement. Frequency features exhibit a marginal change (0.082 to 0.081), suggesting their sensitivity to broader spectral ranges makes them less affected by overlapping speakers. In contrast, the deep speech embedding WavLM shows no consistent improvement (0.080 to 0.081), possibly because the model's learned representation already encodes speaker-invariant information, reducing the relative benefit of explicit diarization.

Table 4.8: Evaluation of video features with and without speaker diarization

Category	Feature	W/O Speaker Diarization	Speaker Diarization
Facial Behaviors	Facial Landmarks	0.058	**0.056**
	Eye Gaze	0.080	**0.077**
	Head Pose	0.074	**0.072**
	Facial Action Units	0.080	**0.078**
Body Pose	Body Keypoints	0.081	**0.079**
Frame Embedding	ViT Embedding	0.084	**0.084**

Video Modality. Table 4.8 shows the corresponding results for visual features. Speaker diarization produces a similar but slightly stronger effect in the video modality, particularly for facial cues. The Facial Landmark features improve from 0.058 to 0.056, while Head Pose and Facial Action Units decrease from 0.074 and 0.080 to 0.072 and 0.078, respectively. These results suggest that clearly segmenting frames belonging to the active listener helps the model focus on relevant head and facial movements instead of those of the current speaker. Eye Gaze also benefits modestly (0.080 to 0.077), indicating that gaze direction is more reliably interpreted once listener frames are isolated. Body Keypoints and ViT embeddings show minimal or no change, reflecting that broad body posture and global frame features are less sensitive to speaker segmentation.

4.6 Discussion

This section synthesises the experimental findings presented above to interpret how different modalities and processing strategies contribute to estimating agreement intensity. In contrast to Chapter 3, which examined when listeners produce backchannel responses, the present discussion focuses on how the expressive properties of those responses reveal the listener's degree of attitudinal alignment.

Unimodal features. The unimodal analyses highlight distinct behavioural signatures captured by audio and video features. Within the audio modality, prosodic descriptors—particularly energy and spectral features—emerge as the most discriminative indicators of agreement. Their effectiveness stems from encoding subtle modulations in loudness, resonance, and smoothness of vocal tone that accompany affirmative utterances such as "yeah" or "right." In contrast, frequency-related cues contribute less, reflecting that pitch contour alone does not fully differentiate agreement from other listener states. Learned embeddings from WavLM achieve performance comparable to handcrafted descriptors, suggesting that both low- and high-level acoustic

representations encode overlapping affective information. As to the visual modality, the strongest signals arise from facial landmarks, head pose, and facial action units, which together capture fine-grained micro-movements such as nodding, eyebrow raising, and smiling. Temporal differencing proves critical for these features, substantially lowering prediction error by emphasising motion-based cues rather than static appearance. Body keypoints and holistic ViT embeddings contribute little, confirming that agreement is primarily conveyed through localized facial dynamics rather than full-body posture or global visual context.

Multimodal fusion. The intra-modality and cross-modality fusion results demonstrate that agreement perception depends on the concurrent expression of complementary acoustic and visual cues. Selective feature integration within each modality—such as combining energy and spectral features for audio, or facial landmarks, head pose, and AUs for video—achieves higher accuracy than fusing all sources indiscriminately. This indicates that redundant or weakly correlated features can obscure the most informative patterns. When both modalities are integrated, the model attains the lowest validation error, underscoring the inherently multimodal nature of agreement behaviour. Audio contributes prosodic warmth and rhythm, while visual cues reveal attentional engagement and affective alignment. Temporal attention analysis further shows that the model dynamically alternates its reliance between modalities: audio weights increase during voiced affirmations, whereas visual cues dominate during silent nods or smiles. Such adaptive weighting reflects the contextual flexibility required to interpret multimodal listener feedback in spontaneous interaction.

Effect of temporal context and preprocessing. To statistically examine the impact of feature preprocessing on regression performance, a paired t-test was conducted to compare the *Subtraction*-based and original representations across temporal windows. As shown in Table 4.9, all comparisons yield highly significant results ($p < 10^{-20}$), confirming that the subtraction method consistently improves both the explained variance and the temporal agreement with ground truth annotations. The

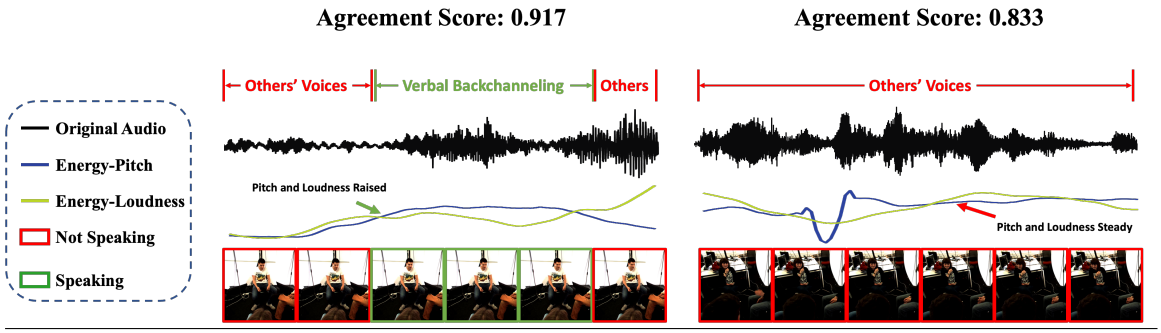
Table 4.9: Paired t-test results comparing different feature engineering methods across observation windows on the MPIIGroupInteraction dataset.

Window	t	p	$R^2_{\mathbf{ori}}$	$R^2_{\mathbf{sub}}$	$\mathbf{CCC_{ori}}$	$\mathbf{CCC_{sub}}$
1 s	-10.13	2.49×10^{-23}	0.036	0.253	0.074	0.368
2 s	-13.42	9.42×10^{-39}	0.040	0.325	0.077	0.450
3 s	-13.61	9.14×10^{-40}	0.036	0.338	0.062	0.487
4 s	-14.27	2.63×10^{-43}	0.024	0.320	0.035	0.448
5 s	-12.57	1.82×10^{-34}	0.035	0.310	0.062	0.442

increase in R^2 values—from approximately 0.03 to over 0.33—indicates that motion-differential preprocessing captures substantially more meaningful behavioural variance than static representations. Likewise, the concordance correlation coefficient (CCC) rises from below 0.08 to nearly 0.49, reflecting a marked improvement in the consistency between predicted and annotated agreement trajectories. Performance peaks at the three-second window, beyond which both R^2 and CCC begin to decline slightly. This suggests that agreement-related cues are primarily concentrated within a short temporal interval preceding the listener's feedback, while longer segments introduce irrelevant context or overlapping behaviours. These findings align with the broader pattern observed in Chapter 3, where moderate contextual spans also yielded the most stable improvements. Collectively, the statistical evidence demonstrates that motion-differential preprocessing enhances sensitivity to attitudinal variations by amplifying localised behavioural dynamics, providing a statistically robust foundation for modelling continuous agreement intensity.

Speaker diarization. The quantitative results in Tables 4.7 and 4.8 indicate that speaker diarization improves model performance across both audio and visual modalities, with the largest relative gain observed in energy-based acoustic features (from 0.079 to 0.076). This improvement suggests that separating speaker roles most effectively clarifies the distribution of vocal energy, which is easily confounded when

Figure 4.5 Visualization of acoustic patterns when the target responds with verbal backchannels versus remains silent.



multiple participants speak simultaneously. To further understand this phenomenon, a qualitative case study was conducted, as illustrated in Figure 4.5.

Figure 4.5 presents two conversational segments annotated with similar agreement levels but exhibiting distinct acoustic–visual characteristics. In the *speaker* condition (left), the target produces a short verbal backchannel (e.g., "yeah") that overlaps with the main speaker's utterance, producing a noticeable rise in both pitch and loudness. In contrast, in the *non-speaker* condition (right), the target remains silent while the main speaker continues speaking; the perceived agreement is instead conveyed through nonverbal gestures such as a nod or brief facial movement. Although the annotated agreement intensity is comparable between the two cases, their acoustic profiles differ substantially: the first involves superimposed vocal activity, while the second contains only the dominant speaker's voice. This comparison highlights how overlapping speech can distort acoustic indicators such as energy or loudness if diarization is not applied. Without role separation, the system may incorrectly attribute the main speaker's high vocal energy to the listener, leading to inaccurate estimation of agreement. Speaker diarization effectively mitigates this issue by isolating listener-specific segments, ensuring that prosodic fluctuations are attributed to the correct individual. The improvement in energy-based and spectral features thus stems from their restored correspondence with true listener activity, rather than contamination from concurrent speech. The same principle extends to the visual

modality, where diarization clarifies which participant's facial or head movements correspond to feedback behaviour. Overall, the case study confirms that accurate speaker segmentation not only refines acoustic feature reliability but also enhances interpretability in multimodal analysis. By isolating listener-specific moments, diarization enables the framework to disentangle overlapping behavioural streams and more faithfully capture the cues that underpin agreement perception.

Limitations regarding cultural context and generalisability. The agreement estimation experiments rely exclusively on the MPIIGroupInteraction dataset, which comprises German-speaking university students in a specific age range and cultural context. Research in cross-cultural pragmatics has established that the expression and perception of agreement are shaped by cultural norms around politeness, directness, and social hierarchy. For instance, cultures that emphasise indirect communication may express agreement through more subtle or ambiguous cues, while those that value direct expression may produce more overt verbal and nonverbal signals. The patterns learned by the proposed framework—such as the relative importance of head nodding versus facial action units—may therefore reflect culturally specific behavioural norms rather than universal indicators of agreement. Although the model achieves strong performance within the studied dataset, its generalisability to other linguistic and cultural contexts remains an open empirical question that warrants investigation in future work.

4.7 Summary

This chapter has extended the investigation of listener feedback from the detection of backchannel occurrences to the estimation of their attitudinal orientation. By focusing on how multimodal backchannel behaviours express different degrees of agreement, the study has provided both behavioural and computational insights into the mechanisms through which social alignment is conveyed in conversation. Compre-

hensive unimodal analyses revealed that energy- and spectral-based acoustic features and fine-grained facial dynamics constitute the most informative cues for agreement estimation. Whereas the acoustic stream captures prosodic smoothness and vocal intensity associated with affirmative responses, visual descriptors such as facial landmarks, head pose, and action units encode subtle nonverbal correlates of agreement, including nods and smiles. Temporal differencing was found to be crucial for video features, enhancing sensitivity to micro-expressions and motion-based signals, while handcrafted audio descriptors remained more stable in their raw form. Through selective intra- and cross-modality fusion, the framework effectively integrated complementary acoustic and visual information, achieving the lowest estimation error among all evaluated configurations. The fusion analyses confirmed that agreement is inherently multimodal: audio cues contribute to the affective tone of verbal acknowledgements, whereas visual cues reveal the listener's attentional and emotional alignment during silent responses. Speaker diarization further refined the estimation by isolating listener-specific segments, reducing interference from active speakers, and enhancing the interpretability of multimodal cues.

Overall, the findings demonstrate that the degree of agreement can be reliably inferred from short, contextually coherent windows of multimodal behaviour. In contrast to the previous chapter, which addressed the temporal occurrence of backchannels, the present work establishes that their expressive form reflects graded social alignment and shared understanding. These results underscore that backchannels are not merely signals of attention but carry meaningful affective and cognitive information about the listener's stance. The next chapter builds on this foundation to explore how such behavioural indicators of agreement and engagement can be leveraged to assess group interaction quality and learning-related outcomes.

Chapter 5

Predicting Learning Gains in Online Service-Learning Programs

5.1 Introduction

Advances in digital technology have substantially reduced the constraints of distance and time in higher education, enabling students to participate in live discussions, collaborate on problem-solving, and share resources within stable online environments. As a result, e-learning has become an integral component of institutional strategy and is widely promoted as a means to broaden participation and enhance instructional quality [40]. Among available platforms, videoconferencing systems such as *Zoom* have become central to synchronous online learning, supporting real-time communication and flexible content delivery [74]. However, this form of instruction also presents challenges. Instructors often face difficulty in observing learners' engagement and understanding through limited on-screen cues. When direct indicators of learning progress are absent, it becomes difficult to adapt teaching strategies in a timely and targeted manner. This challenge has motivated increasing interest in predicting learning gains from the digital traces generated during online sessions. These traces

offer valuable evidence of learner behaviour and interaction, potentially informing adaptive instruction and the design of data-driven pedagogical interventions. Nevertheless, identifying which behavioural signals most effectively represent learning remains difficult because learning processes are multifaceted, temporally dynamic, and context dependent.

Prior research suggests that the multimodal data streams available in videoconferencing environments—spoken language, audio signals, shared-screen content, and video of participants—each provide distinctive insights into cognitive and social dimensions of learning [49, 9]. Multimodal Learning Analytics (MMLA) offers a principled framework for modelling these heterogeneous data sources to capture meaningful behavioural patterns. Despite its potential, two key limitations remain. First, many studies rely primarily on self-report measures rather than automatically recorded behavioural data, limiting ecological validity and objectivity [68, 96]. Second, conventional statistical analyses often fail to exploit the scale and complexity of multimodal datasets, constraining their ability to uncover cross-modal relationships and temporal dynamics [68, 96].

This study addresses these limitations through a multimodal analysis of authentic synchronous service-learning sessions. The dataset comprises four synchronised streams: (1) high-quality audio recordings, (2) automatically generated speech transcripts, (3) shared-screen videos, and (4) gallery-view recordings capturing participants' camera feeds. The inclusion of the gallery-view modality enables examination of visual indicators such as camera usage and body-pose dynamics, yielding a more comprehensive representation of learner behaviour. The study systematically compares alternative feature constructions for each modality and evaluates both single-modality models and multimodal fusion approaches for predicting students' learning gains. The analysis is guided by the following research questions:

RQ1: Among transcript-, audio-, screen-, and gallery-view-based features, which

features of specific single modality provides the most accurate prediction of student learning gains in synchronous videoconference classes?

RQ2: To what extent does multimodal fusion enhance predictive performance compared with unimodal approaches?

Four groups of features were constructed in alignment with the corresponding modalities. *Transcript features* include surface linguistic measures (e.g., word counts) and sentence-level semantic embeddings generated by Sentence Transformers [89]. *Audio features* consist of acoustic descriptors (e.g., loudness, pitch, and Mel-frequency cepstral coefficients) together with contextual paralinguistic representations derived from wav2vec 2.0 [7]. *Screen-content features* are obtained from shared-screen recordings using an automated captioning pipeline for textual content and key-frame semantic representations generated by BLIP-2 [55]. Finally, *gallery-view features* are extracted from participants' camera streams and include indicators of camera usage as well as motion-based features computed from body poses using ViTPose [101]. Predictive models are trained using both unimodal inputs and multimodal fusion strategies, including early- and late-fusion designs. Comparative analyses demonstrate that an attention-based late-fusion model integrating transcript, audio, screen, and gallery-view features achieves the highest predictive accuracy. This study makes three principal contributions. First, it presents a systematic multimodal analysis of authentic videoconference-based instruction that incorporates both screen-sharing and gallery-view recordings linked to validated measures of learning gain. Second, it provides a comprehensive evaluation of unimodal and multimodal fusion models across four complementary data modalities. Third, it identifies cross-modal interaction patterns—spanning linguistic, acoustic, visual, and instructional content—that are most closely associated with positive learning outcomes, offering empirical insights for the design of data-informed, adaptive online teaching.

The remainder of this chapter is organized as follows. Section 3.1 describes the mul-

timodal dataset collected from synchronous service-learning programs, including the session structure and the measurement of self-reported learning outcomes. Section 5.3 details the extraction of linguistic, acoustic, screen-based, and visual features aligned at the utterance level. Section 5.4 presents the experimental design and reports the results for both unimodal and multimodal models. Section 5.5 discusses the findings and their implications. Finally, Section 5.6 summarizes the key contributions and highlights the significance of this study for multimodal learning analytics in online education.

5.2 Dataset

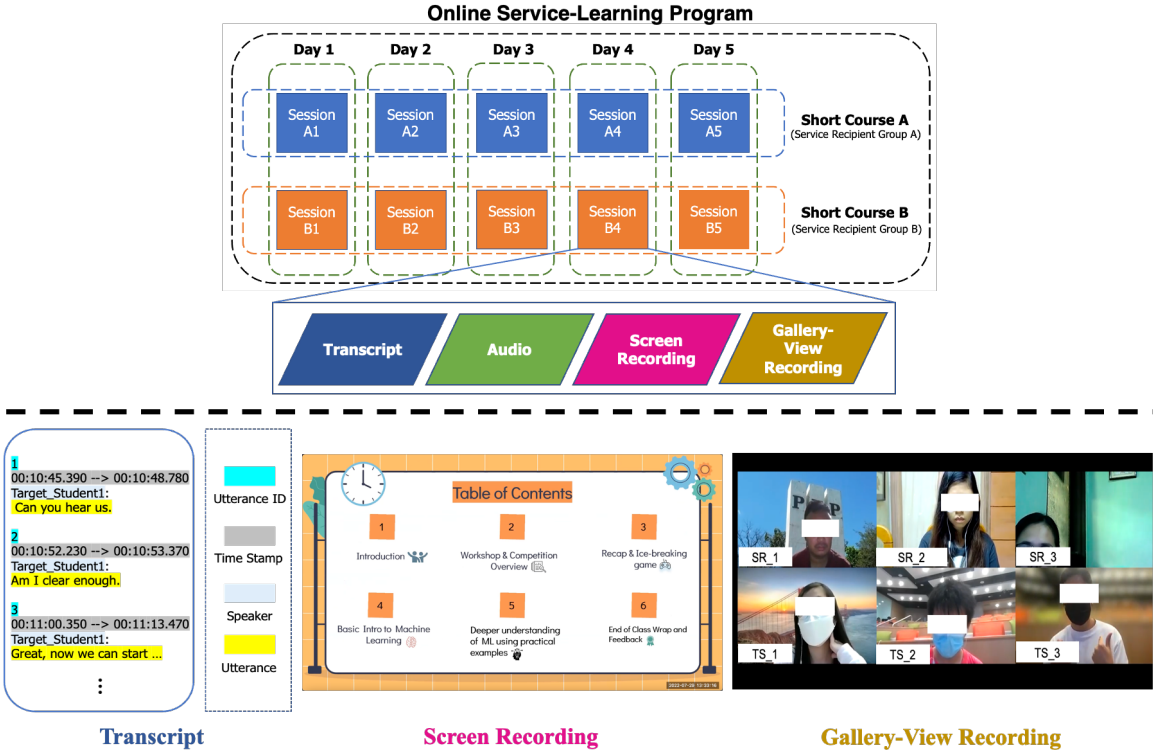
This section describes the context of the online service-learning programs, the multimodal dataset collected from synchronous sessions, and the self-report instrument used to assess student learning gains.

5.2.1 Zoom Dataset for Online Service-Learning Programs

Data for this study were collected from a series of online service-learning programs organised by a public university in Hong Kong during the 2022–2023 academic year, when remote teaching was still widely adopted due to the COVID-19 pandemic. These programs formed part of the university's broader initiative to promote international collaboration through technology-mediated service-learning.

University students enrolled in service-learning courses were organised into small instructional teams responsible for designing and delivering interactive online workshops. The workshops aimed to teach fundamental concepts in artificial intelligence (AI), object recognition, machine learning, and block-based programming. Each tutor team developed its own teaching materials that combined short lectures, live demonstrations, and hands-on exercises designed for accessibility to younger learners. The

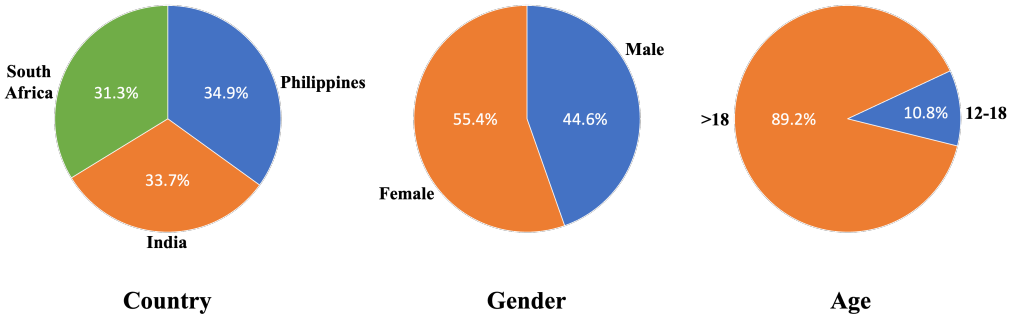
Figure 5.1 Illustration of the data structure and recording examples of the online service-learning programs



workshops were delivered to international service recipients (SRs) located in India, the Philippines, and South Africa, thereby fostering cross-cultural communication and technology exchange within an authentic educational setting.

A total of 166 SRs participated in the study, with an approximately balanced distribution across the three countries, as shown in Figure 5.2. Among them, 44.6% identified as male and 55.4% as female. Most participants were secondary school students aged between 12 and 18, while 10.8% were over 18 years old. On the tutor side, 95 Hong Kong university students participated as instructors. Each program was facilitated by a team of three to four tutors who jointly conducted two five-day workshop sessions. Both sessions followed the same curriculum but were delivered to distinct SR groups, each comprising three to five participants. In total, 25 programs were conducted, producing approximately 250 recorded Zoom sessions for subsequent

Figure 5.2 Statistics of the service-recipients



analysis.

All sessions were hosted and recorded via the Zoom platform, which automatically generated synchronised data streams, including (1) screen-sharing videos, (2) gallery-view recordings, (3) high-quality audio tracks, and (4) time-aligned transcripts. Together, these records provided a comprehensive representation of the online learning environment, encompassing both the instructional content and the interpersonal dynamics between tutors and SRs. Figure 5.1 illustrates the multimodal composition of the dataset.

Each transcript consists of a sequential set of *utterance blocks*, with the following components:

- **Utterance ID:** a unique identifier assigned to each utterance in chronological order;
- **Timestamp:** start and end times recorded in the *hour:minute:second.millisecond* format;
- **Speaker:** an anonymised label, where university tutors are denoted as *Target_Student*{*ID*} (e.g., *Target_Student1*) and service recipients as *Service_Recipient*{*ID*} (e.g., *Service_Recipient1*);
- **Utterance:** the text content of each spoken segment automatically transcribed by Zoom.

The transcript serves as a natural temporal anchor for multimodal alignment and analysis. Each utterance defines a discrete interactional unit that determines the corresponding time window for extracting audio, visual, and screen features. Using these utterance boundaries, all modalities are segmented and synchronised according to the start and end times of each spoken segment. This design enables multimodal features to be aligned at the utterance level. Consequently, the transcript provides a consistent and interpretable reference structure for integrating heterogeneous data streams in subsequent predictive modelling.

5.2.2 Self-Reported Learning Gains Dataset

In this study, we aim at evaluating the learning gains of the university students who served as tutors in the service-learning programs using the multimodal features extracted from the online synchronous session. To evaluate their perceived learning gains, the 95 student tutors were invited to complete a bilingual Chinese–English Self-Reported Learning Gains Inventory (SRLGI) at the conclusion of the courses. The SRLGI was specifically designed to capture their reflections on multiple dimensions of learning outcomes associated with service-learning.

The SRLGI was adapted and validated for this research through a rigorous multi-stage process. Drawing upon prior studies on service-learning assessment, the instrument was tailored to the specific context of these programs. A panel of experienced service-learning educators and researchers reviewed the items to ensure both content and face validity. Subsequent exploratory and confirmatory factor analyses confirmed the internal structure of the four-factor model, with satisfactory fit indices (CFI = 0.973, TLI = 0.956, NFI = 0.971, RMSEA = 0.073), demonstrating the instrument's reliability and construct validity. All participants were informed of the study's purpose and assured that participation was voluntary and unrelated to academic evaluation. Ethical approval for data collection and analysis was granted by the university's research

ethics committee.

Table 5.1: Self-Reported Learning Gains Inventory (SRLGI)

Learning Outcomes	**Items:** *To what extent do you think the service-learning course/program increased or improved your...*
Intellectual	Ability to apply knowledge and skill in real life
	Ability to solve problems
	Ability to think creatively
	Ability to reflect on and learn from your experiences
	Ability to analyze issues from multiple perspectives
Social	Ability to establish and maintain good relationships with other people
	Ability to work with others in a team to achieve common goals
	Respect for people with different backgrounds or perspectives
Civic	Understanding of the needs, potentials, and resources of the community that you served
	Empathy for disadvantaged people
	Commitment to creating a better society
Intrapersonal	Self-confidence
	Self-understanding
	Commitment to continued self-improvement

As shown in Table 5.1, the inventory consists of fourteen items grouped into four learning-outcome dimensions: Intellectual (ILO), Social (SLO), Civic (CLO), and Intrapersonal (IPLO). Each item was rated on a ten-point Likert scale, where 1 indicated "very little" benefit and 10 indicated "very much" benefit. Of the 95 tutors, 66 voluntarily completed the post-program inventory, yielding a response rate of 69.5%. Every service-learning program was represented by at least one response, ensuring complete program-level coverage despite the partial participation rate. For each learning dimension, the mean of its corresponding items was computed to obtain

an overall score.

Figure 5.3 Distribution of averaged learning outcomes: ILO (mean = 7.60, SD = 1.96), SLO (mean = 8.01, SD = 1.81), CLO (mean = 7.87, SD = 1.91), and IPLO (mean = 7.86, SD = 1.88)

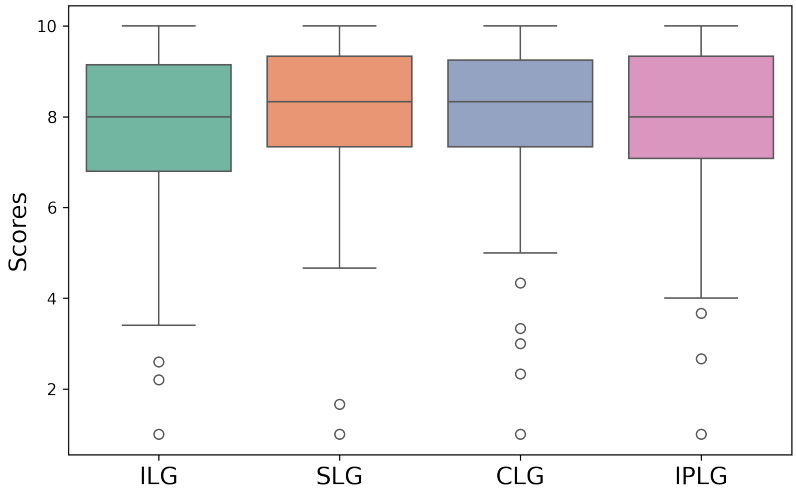


Figure 5.4 Illustration of learning gain classification scheme

	ILO	SLO	CLO	IPLO
Target_Student1	Top 25%	Top 25%	Top 25%	Top 25%
Target_Student2	Middle 50%	Middle 50%	Low 25%	Middle 25%
Target_Student3	Low 25%	Top 25%	Middle 50%	Low 25%
.....	.....	.....	.....	.....
Target_Student66	Top 25%	Top 25%	Low 25%	Middle 25%

Figure 5.3 presents the distribution of the averaged SRLGI scores across the four learning-outcome dimensions: Intellectual (ILO), Social (SLO), Civic (CLO), and Intrapersonal (IPLO). All four dimensions exhibit positively skewed distributions with median values clustered between 7.5 and 8.0, indicating that most tutors reported moderate to high perceived learning gains. The relatively narrow interquartile ranges and standard deviations (approximately 1.8–2.0) suggest a consistent learning ex-

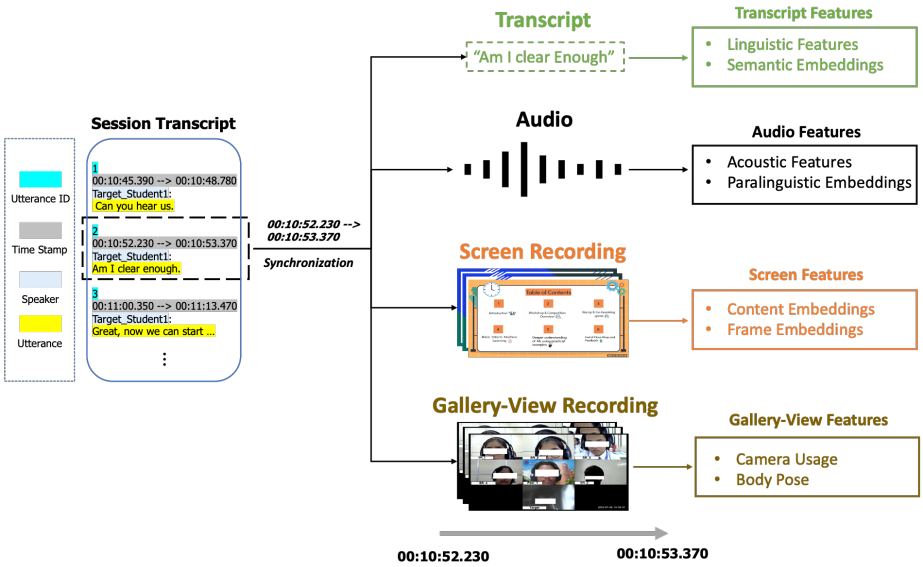
perience across participants, with only a few low-scoring outliers. Among the four categories, the Social and Civic outcomes show slightly higher medians and tighter spreads, implying stronger and more homogeneous development in interpersonal and community-oriented skills. By contrast, the Intellectual and Intrapersonal outcomes display wider score ranges, reflecting greater individual variability in cognitive growth and self-reflective learning.

Task definition: This study aims to predict individual tutors' self-reported learning outcomes using multimodal features extracted from the recorded service-learning sessions. For each tutor, the averaged scores from the Self-Reported Learning Gains Inventory (SRLGI) serve as the target variables corresponding to the four outcome dimensions. To formulate a supervised classification task, the continuous scores are discretised into three levels—Top 25%, Middle 50%, and Bottom 25%—according to their empirical distribution (Figure 5.4). This process yields 16, 34, and 16 samples in the top, middle, and bottom categories, respectively. The predictive models take as input the temporally aligned transcript, audio, and screen-based features described in the previous sections. This formulation enables systematic examination of how multimodal behavioural patterns correspond to different levels of self-perceived learning gains and evaluates the relative contribution of linguistic, acoustic, and visual modalities. By combining session-level multimodal data with post-program self-assessments, the study provides an integrated view of how tutors' interactional and instructional behaviours are associated with their learning development in online service-learning environments.

5.3 Feature Extraction

As illustrated in Figure 5.5, the *utterance* serves as the fundamental unit for synchronizing multimodal feature extraction from the transcript, audio, and screen-sharing streams. Each utterance is precisely defined by its *timestamp* in the transcript, pro-

Figure 5.5 Overall framework for multimodal feature extraction, illustrating the alignment of transcript, audio, and screen-sharing data at the utterance level and the corresponding feature groups derived from each modality



viding an explicit start and end time for alignment across modalities. For example, an utterance spanning *00:10:52.230–00:10:53.370* specifies the exact temporal window for which the corresponding acoustic and visual features are extracted. This design ensures that all features correspond to the same communicative event, allowing the analysis to capture not only the linguistic content of what was said but also the accompanying vocal and visual expressions. The use of the utterance as the temporal reference unit offers several advantages. First, it provides a natural segmentation of interaction that reflects conversational flow rather than arbitrary time slicing. Second, it enables consistent multimodal alignment by matching the audio waveform, screen content, and camera-view frames directly to the linguistic transcript without resampling or interpolation. Through this alignment scheme, multimodal data streams are converted into utterance-level feature vectors that can be integrated into downstream predictive models.

Transcript Features

Linguistic characteristics of spoken utterances provide important cues for understanding engagement, cognition, and communication processes in learning environments [1, 58]. In this study, both statistical and semantic features were extracted from each transcribed utterance, covering surface-level linguistic indicators and higher-level semantic representations. The extracted features are summarised as follows:

1. **Basic Linguistic Features:** For each utterance, we compute the total word count, the count of content words (nouns, verbs, adjectives, and adverbs), and the utterance duration derived from the transcript timestamps. These measures reflect linguistic productivity and fluency, which are often associated with the speaker's communicative involvement and cognitive effort.

2. **LIWC-Based Psycholinguistic Features:** The Linguistic Inquiry and Word Count (LIWC) lexicon categorises words into psychologically meaningful dimensions, enabling the analysis of cognitive, emotional, and social aspects of language use. In this work, we focus on two major categories: *cognitive processes* (e.g., insight, causation) and *affective states* (e.g., positive and negative emotions). Cognitive-process terms reflect reasoning and reflection, while affective-state terms capture emotional tone and motivational stance during communication.

3. **Dialogue Act Features:** Each utterance is annotated with one or more dialogue act categories to represent its communicative function, including Greeting, Question, Answer, Feedback, Agreement, Disagreement, Clarification, Request, Suggestion, Thanking, and Closing. Because Zoom transcriptions often include compound or multifunctional sentences, multiple dialogue act labels may co-occur within a single utterance. These annotations provide a pragmatic layer that complements linguistic content with interactional intent.

4. **Semantic Embeddings:** To capture the contextual meaning beyond surface-level word use, each utterance is encoded into a 768-dimensional vector using a pre-trained Sentence-BERT model [81]. These embeddings represent semantic similarity and discourse coherence across utterances in a continuous vector space, enabling downstream models to learn patterns of meaning and topic alignment that are not explicitly observable in lexical statistics.

Audio Features

Acoustic features extracted from speech provide valuable information for analysing engagement, affective states, and interaction quality in educational contexts [87]. Beyond basic signal descriptors, paralinguistic characteristics such as prosody and speaking style convey subtle cues about communicative behaviour and emotional involvement [72]. In this study, four groups of audio features are extracted for each utterance, combining low-level acoustic descriptors from the eGeMAPS feature set with high-level contextual embeddings derived from a self-supervised speech model.

1. **Frequency-Related Features:** Capture pitch and harmonic properties of speech, including measures such as fundamental frequency (F0), formant frequencies, jitter, and harmonic-to-noise ratio. These features provide information about pitch variability and prosodic patterns during interaction.

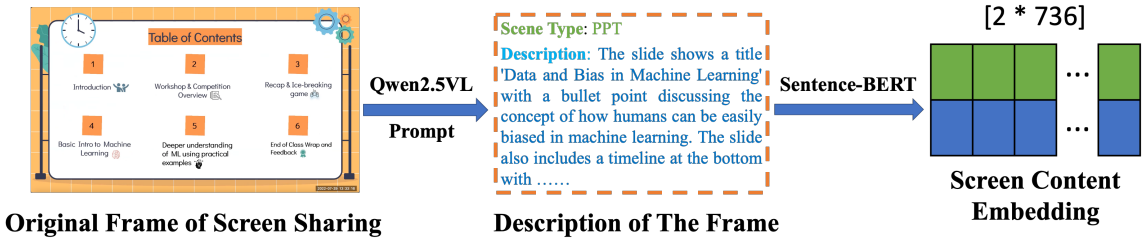
2. **Energy-Related Features:** Describe the loudness and energy dynamics of speech, including overall loudness, standard deviation, and the range and slope of energy changes. Such features reflect vocal intensity and emphasis, which are linked to engagement and affective expression.

3. **Spectral-Related Features:** Characterise the spectral shape and timbre of speech, with measures such as spectral flux, Mel-frequency cepstral coefficients (MFCCs), and spectral slopes. These features capture the richness and clarity

of the voice signal and are informative for distinguishing variation in speech quality.

4. **Paralinguistic Embeddings:** High-level representations derived from a pre-trained wav2vec 2.0 model [7]. Each utterance is encoded as a fixed-length (768) embedding that captures complex prosodic and contextual acoustic patterns beyond those represented by hand-crafted descriptors.

Figure 5.6 Illustration of screen content embedding extraction



Screen Recording Features

The quality and delivery of instructional content, conveyed through screen presentations in online teaching, are recognised as significant factors influencing student engagement and learning outcomes [49, 9]. Visual information, alongside spoken language and audio cues, shapes how learners process and retain material. To capture both the informational content and the visual characteristics of screen sharing, we extract two types of features from the video recordings:

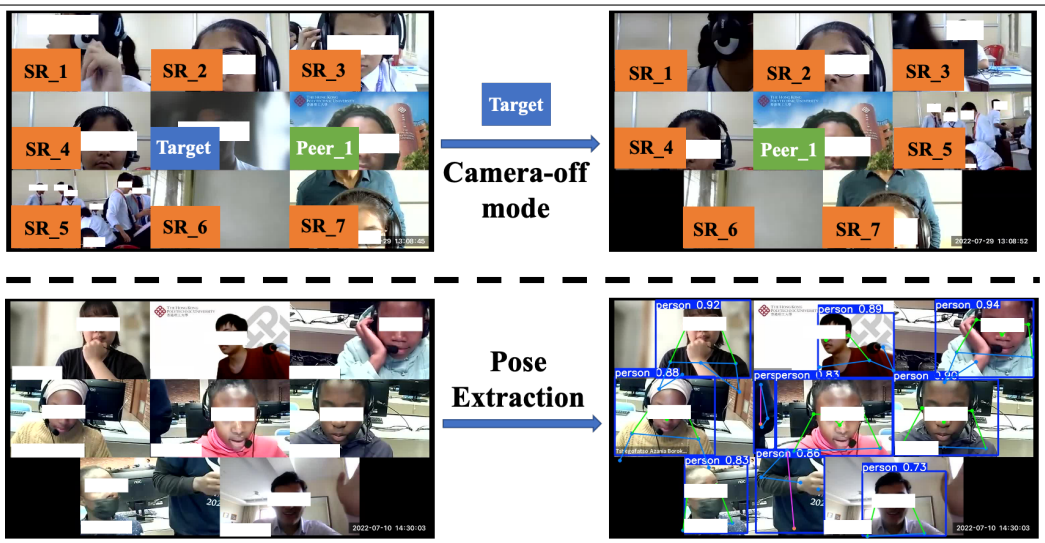
1. **Content Embeddings:** Each keyframe from the shared-screen recordings is first processed using the multimodal vision–language model Qwen2.5-VL [8], which automatically identifies the *scene type* (e.g., slide, video, coding interface) and generates a concise *description* of the on-screen content, such as titles, bullet points, or visual elements. The combined textual output—comprising both

scene type and content description—is then encoded into semantic embeddings using the same Sentence-BERT model employed for transcript features. This design ensures cross-modality consistency and allows the semantic alignment of spoken and visual instructional content within a unified embedding space.

2. **Keyframe Embeddings:** To complement the textual description, direct visual embeddings are extracted from the same keyframes using BLIP-2 [55], a vision–language pre-trained model capable of encoding high-level scene semantics. These embeddings capture spatial layout, visual complexity, and stylistic structure, providing information about how instructional content is visually organised and presented.

By integrating textual and visual representations, this feature set captures both the conceptual meaning of instructional materials and the perceptual characteristics of their delivery, offering a holistic view of how screen-shared content supports communication and learning.

Figure 5.7 Examples of camera usage change and pose extraction



Gallery-View Recording Features

The gallery-view recordings capture the simultaneous video feeds of all participants during the synchronous sessions, providing visual information about their on-camera presence and nonverbal behaviour. As illustrated in Figure 5.7, each frame contains multiple participants—tutors, peers, and service recipients (SRs)—whose video streams are analysed to extract two categories of features: *camera usage* and *body pose*.

1. **Camera Usage:** Prior studies have shown that camera usage in videoconferencing can influence learners' satisfaction, sense of presence, and perceived learning outcomes [15, 47, 9]. In this study, camera-usage features are computed *per utterance* for the target tutor, aligned to the utterance's start–end timestamps. Let u denote an utterance and F_u the number of gallery-view frames that fall within the utterance window. Define $C_{u,f} \in \{0, 1\}$ as an indicator that equals 1 if the target's camera is on in frame f of utterance u, and 0 otherwise. Then,

$$\mathrm{CCT}_u = \sum_{f=1}^{F_u} \mathbb{I}(C_{u,f} = 1), \qquad \mathrm{CRT}_u = \frac{1}{F_u} \sum_{f=1}^{F_u} \mathbb{I}(C_{u,f} = 1),$$

   where CCT_u represents the camera-on frame count within utterance u, and CRT_u denotes the corresponding proportion of on-camera frames within that utterance. When aggregated beyond the utterance level—such as by role condition or session date—these utterance-level statistics are averaged accordingly.

2. **Body Pose:** Body-pose features are extracted from gallery-view frames using ViTPose [101], which detects skeletal keypoints of each participant in real time. From these keypoints, two motion-based descriptors are computed: (1) the *Pose of Target (PT)*, representing the aggregated body-pose features of the target tutor across frames, which describe overall posture and movement dynamics during interaction; and (2) the *Body Movement of Target (BMT)*, representing

the frame-wise variation of the target's body pose, measured as the temporal difference between consecutive pose vectors, as introduced in the feature engineering procedure in Chapter 3. BMT reflects the magnitude of movement change and serves as an indicator of physical activity or visible engagement intensity captured through the camera.

To contextualise these visual features, all gallery-view recordings are segmented according to the active speaker identified from the transcript. Each segment is labelled as one of three conversational conditions: (1) *Target Speaking*, when the focal tutor is speaking; (2) *Peer Speaking*, when another tutor is speaking; and (3) *Service Recipient Speaking*, when an SR is speaking. Within each condition, camera-usage and body-pose features are averaged to represent the participants' overall visual state. This role-based contextualisation enables the analysis of how tutors and learners maintain visual presence and bodily activity across different phases of group interaction, linking observable nonverbal behaviour to learning dynamics in online service-learning sessions.

5.4 Experiment Results

5.4.1 Experiment Design

To evaluate the effectiveness of different modelling strategies for multimodal feature analysis, we systematically compare alternative approaches to feature construction and fusion. Multimodal features are extracted at the utterance level for each target student, and two methods are employed to construct the final feature representation for prediction: *utterance-based* and *date-based*.

In the *utterance-based* method, features from all utterances spoken by a target student are concatenated in temporal order to form a feature matrix of size $[N_{\text{utterance}}, D_{\text{feat}}]$,

where $N_{\text{utterance}}$ is the number of utterances and D_{feat} is the dimensionality of each feature vector. This sequence is then modelled with a gated recurrent unit (GRU) to predict the student's learning outcomes. In the *date-based* method, utterances are grouped by session date, producing feature matrices of size $[N_{\text{date}}, D_{\text{feat}}]$, where N_{date} denotes the number of days in which the student participated. GRU models are then applied to capture temporal dependencies across days. Both feature construction methods are implemented under unimodal and multimodal conditions. For multimodal experiments, we compare two fusion strategies: (1) direct concatenation, which merges feature vectors from different modalities, and (2) attention-based fusion, which dynamically assigns weights to each modality according to its predictive contribution.

All models are trained and evaluated using student-level five-fold cross-validation, with results reported as the average across folds. Early stopping with a patience of 10 epochs is applied based on validation loss, with a maximum of 200 training epochs per run. Optimization is performed using Adam with a learning rate of 1×10^{-4}. Hyperparameters are selected through heuristic search to identify appropriate configurations for each experimental condition. The dataset of 66 target students is randomly divided into a training set of 49 and a test set of 17, corresponding to a 0.75 train-test split. Model performance is assessed by predicting self-reported learning outcomes using the constructed feature representations.

5.4.2 Unimodal Features

The unimodal experiments assess the predictive power of individual feature sets derived from transcript, audio, screen, and gallery-view modalities. Results are compared across feature types and feature construction strategies. Model performance is evaluated for four self-reported learning gains categories: *Intellectual*, *Social*, *Civic*, and *Intrapersonal*. Each modality is modelled independently to identify which fea-

Table 5.2: Predictive performance of **transcript** features under two feature construction methods

Features	Feature Construction	Intellectual		Social		Civic		Intrapersonal	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Speech Features	Utterance	0.53	0.51	0.55	0.54	0.42	0.41	0.64	0.62
	Date	**0.59**	**0.58**	**0.57**	**0.56**	**0.45**	**0.44**	**0.68**	**0.67**
LIWC Features	Utterance	0.54	0.52	0.58	0.57	0.62	0.61	0.49	0.48
	Date	**0.58**	**0.56**	**0.60**	**0.59**	**0.65**	**0.64**	**0.51**	**0.50**
Dialogue Act	Utterance	0.49	0.48	0.62	0.61	0.46	0.45	0.44	0.42
	Date	**0.55**	**0.54**	**0.66**	**0.65**	**0.50**	**0.49**	**0.47**	**0.46**
Utterance Embeddings	Utterance	0.53	0.57	0.55	0.54	**0.54**	**0.53**	0.48	0.47
	Date	**0.64**	**0.62**	**0.56**	**0.55**	0.52	0.51	**0.64**	**0.63**

tures are most indicative of learning gains and how different feature construction methods influence predictive effectiveness. Specifically, transcript features capture linguistic and semantic patterns in tutors' speech; audio features model vocal and prosodic characteristics; screen features represent the instructional content being shared; and gallery-view features quantify visible participation and body pose during online interaction. This decomposition enables a systematic examination of how each modality, considered in isolation, relates to tutors' perceived learning development before exploring cross-modal dependencies in subsequent sections.

Transcript Features As shown in Table 5.2, transcript-based models demonstrate solid predictive capability across all four learning outcomes, with performance ranging between 0.42 and 0.68 in weighted accuracy. Among the feature types, *Utterance Embeddings* consistently achieve the strongest overall results, reaching 0.64 in ac-

curacy and 0.62 in F1-score for both the *Intellectual* and *Intrapersonal* outcomes under the date-based configuration. This indicates that high-level semantic representations effectively capture discourse coherence and conceptual depth in tutors' language, which are closely tied to reflective learning and cognitive growth. *LIWC Features* also yield competitive performance, particularly for *Civic* (0.65) and *Social* (0.60) outcomes, outperforming other linguistic categories in these dimensions. This pattern suggests that psychologically informed lexical indicators—such as words associated with reasoning, empathy, and social processes—are reliable markers of community-oriented and collaborative learning. *Dialogue Act Features* show moderate but distinctive predictive power, with notably high performance on *Social* learning (0.66 accuracy, 0.65 F1), reflecting that communicative functions such as giving feedback, agreeing, and responding to peers correlate with the development of interpersonal skills. In contrast, surface-level *Speech Features* perform relatively lower in most dimensions (0.53–0.59 accuracy) but attain the best results for *Intrapersonal* learning (0.68 accuracy, 0.67 F1). This may reflect that verbosity, elaboration, and temporal pacing are linked to tutors' self-reflective expression and confidence during instruction. Across all feature categories, the date-based construction method generally outperforms utterance-based modelling, with improvements of approximately 0.03–0.07 across outcomes. This may indicate that aggregating linguistic behaviour over entire sessions captures more stable communicative patterns, reducing noise from local variations at the utterance level. Overall, transcript-based models reveal that both semantic and psychologically grounded linguistic indicators are robust predictors of tutors' perceived learning gains, particularly for intellectual and social-cognitive dimensions.

Audio Features As presented in Table 5.3, audio-based features display moderate yet consistent predictive performance across all learning outcomes, with accuracies ranging between 0.40 and 0.62. Among the four feature categories, *Paralinguistic*

Table 5.3: Predictive performance of **audio** features under two feature construction methods

Features	Feature Construction	Intellectual		Social		Civic		Intrapersonal	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Frequency	Utterance	0.45	0.43	0.48	0.46	0.41	0.39	0.52	0.50
	Date	**0.49**	**0.47**	**0.50**	**0.49**	**0.44**	**0.42**	**0.55**	**0.54**
Energy	Utterance	0.48	0.46	0.52	0.51	**0.56**	**0.55**	0.44	0.43
	Date	**0.51**	**0.50**	**0.54**	**0.53**	0.46	0.45	**0.59**	**0.58**
Spectral	Utterance	0.44	0.42	0.47	0.45	0.40	0.38	0.50	0.49
	Date	**0.47**	**0.45**	**0.49**	**0.48**	**0.43**	**0.41**	**0.53**	**0.52**
Paralinguistic Embeddings	Utterance	0.56	0.55	0.53	0.51	0.47	0.46	0.58	0.57
	Date	**0.62**	**0.60**	**0.55**	**0.54**	**0.50**	**0.49**	**0.62**	**0.61**

Embeddings achieve the highest overall results, particularly for the *Intellectual* (0.62 accuracy, 0.60 F1) and *Intrapersonal* (0.62 accuracy, 0.61 F1) outcomes under the date-based configuration. These embeddings, derived from *wav2vec 2.0*, encode rich prosodic and contextual acoustic representations that extend beyond handcrafted signal descriptors, capturing subtle variations in speaking style, tone, and emphasis associated with cognitive engagement and self-reflective expression. Low-level descriptors—including *Frequency*-, *Energy*-, and *Spectral*-related features—show complementary yet differentiated predictive patterns. Frequency features exhibit steady performance across all learning dimensions (0.45–0.50 accuracy), suggesting that prosodic variation such as pitch modulation reflects attentional focus and communicative adaptability during instruction. Energy-related features yield higher scores in the *Civic* outcome (0.56 accuracy, 0.55 F1 in utterance-based) and strong performance for *Intrapersonal* learning (0.59 accuracy, 0.58 F1 in date-based), indicating that vocal intensity and loudness dynamics may signal tutors' affective involvement

Table 5.4: Predictive performance of **screen** features under two feature construction methods

Features	Feature Construction	Intellectual		Social		Civic		Intrapersonal	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Content Embeddings	Utterance	0.60	0.59	0.45	0.43	0.50	0.49	0.51	0.50
	Date	**0.67**	**0.66**	**0.48**	**0.46**	**0.54**	**0.53**	**0.55**	**0.54**
Keyframe Embeddings	Utterance	0.50	0.48	0.54	0.53	**0.53**	**0.52**	0.45	0.43
	Date	**0.53**	**0.51**	**0.56**	**0.55**	0.48	0.46	**0.57**	**0.56**

and assertiveness in guiding interactions. Spectral features produce slightly lower but stable results (0.40–0.53 accuracy), which likely capture articulation clarity and sound quality nuances that correlate with communication fluency.

Screen Recording Features As shown in Table 5.4, screen-based features yield competitive predictive performance, revealing that the instructional content displayed through screen sharing provides meaningful cues about tutors' learning outcomes. The *Content Embeddings*, generated from textual descriptions of the shared-screen keyframes, achieve the highest overall accuracy across outcomes, with particularly strong performance in the *Intellectual* dimension (0.67 accuracy, 0.66 F1). This suggests that the semantic richness and topical diversity of the presented material correlate with tutors' cognitive engagement and conceptual mastery. Tutors who deliver more content-dense, well-structured explanations may also demonstrate deeper understanding and reflective learning. By contrast, *Keyframe Embeddings*, which encode the visual composition and layout of the shared screen, show relatively stronger results for the *Social* (0.56 accuracy, 0.55 F1) and *Intrapersonal* (0.57 accuracy, 0.56 F1) categories. These outcomes imply that visually engaging and dynamic instructional materials—such as activity slides, demonstrations, or graphical elements—may foster

Table 5.5: Predictive performance of **gallery-view** features under two feature construction methods

Features	Feature Construction	Intellectual		Social		Civic		Intrapersonal	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Camera-On Count (CCT)	Utterance	**0.55**	**0.53**	0.59	0.57	0.56	0.54	0.27	0.26
	Date	0.54	0.52	**0.63**	**0.61**	**0.60**	**0.58**	**0.30**	**0.29**
Camera-On Ratio (CRT)	Utterance	0.50	0.48	**0.61**	**0.60**	0.54	0.52	0.26	0.25
	Date	**0.52**	**0.51**	0.59	0.57	**0.57**	**0.56**	**0.28**	**0.27**
Pose of Target (PT)	Utterance	**0.58**	**0.56**	0.52	0.51	0.59	0.57	0.33	0.32
	Date	0.56	0.54	**0.55**	**0.53**	**0.58**	**0.57**	**0.34**	**0.33**
Pose-Difference of Target (PDT)	Utterance	0.57	0.55	**0.57**	**0.55**	**0.61**	**0.60**	0.33	0.32
	Date	**0.59**	**0.57**	0.55	0.53	0.59	0.57	**0.35**	**0.34**

interpersonal collaboration and self-reflective awareness. Interestingly, the *Civic* outcome displays mixed results: utterance-based keyframe features slightly outperform date-based ones (0.53 vs. 0.48), suggesting that short-term variations in instructional context or screen transitions may capture momentary social responsiveness that aggregated representations tend to smooth out.

Gallery-view Recording Features As shown in Table 5.5, gallery-view features exhibit consistent predictive capability across learning outcomes, demonstrating that visible participation and nonverbal behaviour captured through the camera provide informative cues for modelling learning-related engagement. Overall, accuracy ranges between 0.25 and 0.63, with the strongest results observed for the *Social* and *Civic* outcomes, reflecting the interactional and community-oriented characteristics of the service-learning context.

For the *camera-usage features*, both the *Camera-On Count (CCT)* and the *Camera-On Ratio (CRT)* achieve moderate but reliable performance. CCT slightly outperforms CRT across most outcomes, reaching 0.63 accuracy and 0.61 F1 for *Social* learning and 0.60 accuracy and 0.58 F1 for *Civic* learning under the date-based configuration. These results suggest that tutors who appear on camera more frequently are more engaged in interaction and more active in maintaining social presence. In contrast, CRT, which reflects the proportion of frames in which the camera remains active, performs marginally lower but still captures complementary aspects of sustained participation and attentiveness. The *body-pose features* provide additional insight into tutors' physical activity and embodied engagement. The *Pose of Target (PT)* features, which represent overall posture and movement distribution, achieve competitive performance across all dimensions, with the highest accuracy of 0.58 for *Intellectual* learning in the utterance-based model. The *Pose-Difference of Target (PDT)* features, capturing frame-to-frame pose variation, yield the best results overall, reaching 0.59 accuracy (0.57 F1) for *Intellectual* and 0.61 accuracy (0.60 F1) for *Civic* outcomes. These findings indicate that dynamic bodily movements—such as head nods, gestural emphasis, and posture shifts—reflect cognitive and affective engagement signals that correlate with perceived learning.

5.4.3 Multimodal Features

To evaluate the joint contribution of linguistic, acoustic, and visual modalities, multimodal experiments were conducted by combining transcript, audio, screen, and gallery-view features under two fusion strategies: concatenation and attention-based fusion. Results are summarised in Table 5.6. Overall, fusing complementary modalities led to consistent performance gains over unimodal baselines, confirming that learners' behaviours and communicative cues are more effectively represented when information from multiple channels is integrated.

Table 5.6: Predictive performance of **multimodal** features under two fusion strategies

Feature Combination	**Fusion Strategy**	**Intellectual**		**Social**		**Civic**		**Intrapersonal**	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Transcript + Audio	Concat	**0.61**	**0.59**	0.70	0.66	0.51	0.49	0.68	0.66
	Attention	0.60	0.58	**0.72**	**0.69**	**0.53**	**0.51**	**0.69**	**0.68**
Transcript + Screen	Concat	0.68	0.66	0.63	0.61	**0.55**	**0.53**	0.67	0.65
	Attention	**0.69**	**0.68**	**0.65**	**0.63**	0.54	0.52	**0.68**	**0.67**
Transcript + Gallery	Concat	0.65	0.63	0.68	0.66	0.55	0.53	0.65	0.64
	Attention	**0.67**	**0.65**	**0.69**	**0.67**	**0.56**	**0.55**	**0.67**	**0.66**
Audio + Screen	Concat	**0.59**	**0.57**	**0.63**	**0.61**	0.48	0.46	**0.62**	**0.61**
	Attention	0.57	0.55	0.62	0.60	**0.49**	**0.47**	0.61	0.60
Audio + Gallery	Concat	0.58	0.56	0.65	0.63	0.52	0.50	0.63	0.62
	Attention	**0.60**	**0.58**	**0.66**	**0.64**	**0.53**	**0.52**	**0.64**	**0.63**
Screen + Gallery	Concat	**0.63**	**0.61**	0.64	0.62	**0.54**	**0.52**	0.65	0.63
	Attention	0.62	0.60	**0.65**	**0.63**	0.53	0.51	**0.66**	**0.65**
Transcript + Audio + Screen	Concat	0.68	0.67	0.71	0.69	0.53	0.52	0.69	0.68
	Attention	**0.71**	**0.70**	**0.74**	**0.72**	**0.55**	**0.54**	**0.72**	**0.71**
Transcript + Audio + Gallery	Concat	0.69	0.67	**0.73**	**0.71**	0.56	0.54	0.70	0.68
	Attention	**0.71**	**0.69**	0.72	0.70	**0.57**	**0.55**	**0.72**	**0.70**
Transcript + Screen + Gallery	Concat	**0.69**	**0.67**	**0.71**	**0.69**	**0.56**	**0.54**	0.70	0.68
	Attention	0.68	0.66	0.70	0.68	0.55	0.53	**0.71**	**0.70**
Audio + Screen + Gallery	Concat	**0.61**	**0.59**	**0.63**	**0.61**	0.49	0.47	0.63	0.61
	Attention	0.60	0.58	0.61	0.59	**0.51**	**0.49**	**0.64**	**0.62**
All (T + A + S + G)	Concat	0.69	0.68	0.72	0.70	0.55	0.54	0.70	0.69
	Attention	**0.73**	**0.72**	**0.75**	**0.73**	**0.57**	**0.56**	**0.73**	**0.72**

Among two-modality combinations, *Transcript+Audio* and *Transcript+Gallery* achieved the strongest results, particularly on the *Social* and *Intrapersonal* outcomes. These improvements suggest that coupling spoken content with vocal and visual responsiveness provides a more complete view of interpersonal engagement and self-expression. In contrast, the *Audio+Screen* and *Audio+Gallery* pairs exhibited modest improvements, indicating that purely perceptual channels without linguistic grounding are less predictive of learning gains. The *Screen+Gallery* combination performed moderately well, implying that observable on-screen activity and visible body movement capture complementary but weaker behavioural signals compared to speech-related features.

For three-modality settings, the inclusion of transcript features consistently enhanced predictive accuracy across outcomes. The combinations *Transcript+Audio+Screen* and *Transcript+Audio+Gallery* both surpassed their bimodal counterparts, with the latter showing particularly high F1-scores on the *Social* dimension (0.73) and *Intrapersonal* learning (0.72). This indicates that integrating verbal, paralinguistic, and visual behaviours helps model not only cognitive but also relational aspects of engagement during collaborative learning.

The four-modality fusion ($T+A+S+G$) achieved the overall best results, reaching an F1-score of 0.73 on *Social* learning and 0.72 on *Intrapersonal* outcomes. However, the gains over three-modality settings were incremental, suggesting a saturation effect where additional modalities contribute diminishing returns once core linguistic and affective signals are included. Furthermore, attention-based fusion was not universally superior—simple concatenation performed comparably or slightly better in several configurations such as *Transcript+Audio* and *Screen+Gallery*. This indicates that, under limited training data, explicit weighting mechanisms do not always generalise better than direct feature integration. Overall, these findings highlight that multimodal modelling benefits from a balanced fusion of language, sound, and vision, where each channel enriches different dimensions of learners' behavioural representa-

tion.

5.5 Discussion

RQ1: Which single-modality feature set (transcript, audio, screen, or gallery-view descriptors) most accurately predicts student learning gains in synchronous service-learning classes? Among the four modalities, transcript features emerge as the most effective predictors of student learning gains. Utterance embeddings derived from Sentence-BERT models capture not only lexical content but also semantic coherence and contextual relevance within instructional dialogue. When aggregated at the session level, these embeddings reflect the degree of conceptual organisation and the tutors' ability to sustain meaningful discourse over time. Their consistent superiority across learning outcomes indicates that language-based representations are closely aligned with cognitive engagement and reflective understanding. LIWC-based psycholinguistic features further enrich this linguistic perspective by quantifying metacognitive and affective markers—such as causal reasoning and reflective verbs—that are theoretically linked to deep learning processes. Dialogue-act features complement these signals by modelling the communicative intent of utterances; acts such as clarification, feedback, and agreement correspond to collaborative sense-making and social coordination, which are particularly relevant to Social and Civic learning outcomes.

Audio features provide complementary but comparatively weaker predictive power. Paralinguistic embeddings from wav2vec 2.0 encode prosodic variation and vocal style, yielding moderate accuracy for the Intellectual and Intrapersonal dimensions. These cues capture how tutors express themselves—through tone, emphasis, and rhythm—rather than what they say, and are thus more indicative of affective involvement and confidence than of cognitive depth. Low-level frequency, energy, and spectral features show more limited contribution, reflecting that surface acoustic variation

alone does not directly signify conceptual understanding or reflective engagement.

Figure 5.8 Average frame counts of different scene types across intellectual gain groups

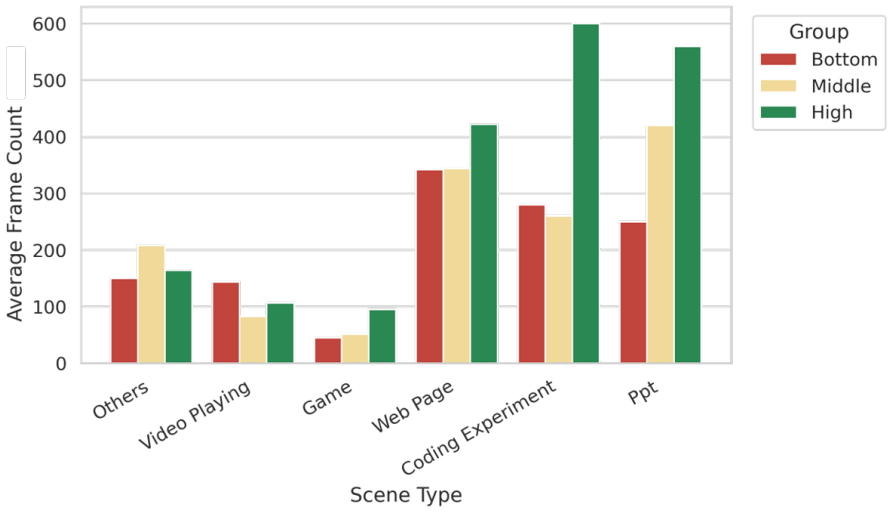


Table 5.7: Scene type statistics under different learning gain groups (ILO).

Scene Type	**Bottom**	**Middle**	**High**	F**-value**	p**-value**
Others	150.000	208.484	164.438	0.310	0.734
Video Playing	143.714	82.935	106.375	1.459	0.241
Game	45.071	51.645	95.313	1.592	0.212
Web Page	341.929	344.258	422.063	0.201	0.819
Coding Experiment	280.145	261.389	607.125	315393.800	**0.000****
PPT	255.348	426.164	563.211	323882.465	**0.000****

Screen-based features demonstrate strong predictive power for the Intellectual learning dimension, highlighting the role of instructional content and presentation context in shaping cognitive outcomes. The analysis of scene-type distributions (Figure 5.8 and Table 5.7) reveals significant differences across learning-gain groups, particularly

for the *Coding Experiment* and *PPT* scenes ($p < 0.001$). High-gain tutors spent substantially more time presenting slides and conducting coding demonstrations, indicating that active engagement with structured and content-rich instructional materials corresponds to higher intellectual learning gains. In contrast, low-gain participants exhibited relatively greater proportions of passive or unstructured scenes such as video playing or browsing, suggesting a weaker instructional focus. The superior performance of content embeddings (0.67 F1) further supports this interpretation. These embeddings, which capture the semantic richness and organisational clarity of shared materials, appear to reflect tutors' conceptual depth and pedagogical preparation more effectively than purely visual or aesthetic features. By contrast, keyframe embeddings show only moderate predictive accuracy, implying that the layout or visual design of materials contributes less to learning outcomes than their semantic content. Overall, these results demonstrate that cognitively active and well-structured screen activities—particularly those involving explanation, demonstration, and synthesis—serve as reliable indicators of intellectual learning gain.

Figure 5.9 Average camera-on frame counts of target students across different speaking roles (Target, Peer, and SR) under varying learning gain groups

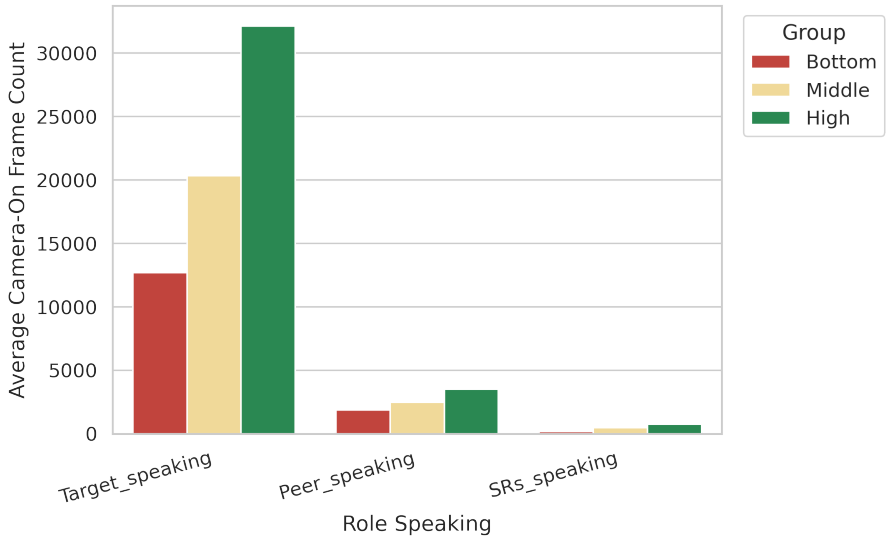


Table 5.8: Camera-on frame statistics of target students across different speaking roles under varying learning gain groups

Role (Speaking to Target)	**Bottom**	**Middle**	**High**	**F-value**	**p-value**
Target Speaking	12,687.38	20,325.88	32,117.25	1,099.39	**0.000****
Peer Speaking	1,874.36	2,487.69	3,512.70	539.83	**0.000****
SR Speaking	192.19	467.50	769.11	1,152.64	**0.000****

The gallery-view analysis extends the multimodal framework to visible social and embodied engagement. As shown in Figure 5.9 and Table 5.8, camera-usage statistics vary significantly across learning-gain groups ($p < 0.001$). High-gain participants maintain substantially higher camera-on frame counts across all speaking roles, with the largest increase observed during their own speaking turns (over 32,000 frames on average). This consistent on-camera presence suggests stronger social engagement and willingness to communicate visually, aligning with the collaborative and civic objectives of the service-learning environment. In addition, higher camera-on rates when peers or service recipients are speaking indicate attentive participation and responsive listening—behaviours that likely contribute to perceived social connectedness and empathy. These findings correspond to the positive correlations observed between camera-usage features (CCT and CRT) and Social and Civic learning gains. Body-pose dynamics, particularly the Pose-Difference of Target (PDT), achieve the strongest Civic and Intellectual predictions (0.61 F1), showing that visible bodily expressiveness and micro-movements serve as secondary but meaningful cues of engagement and reflection. Together, the gallery-view results suggest that sustained visibility, active physical presence, and responsive nonverbal behaviours distinguish more socially and cognitively engaged learners.

RQ2: To what extent does multimodal fusion enhance predictive performance compared with the best single-modality baseline? Multimodal fusion

generally improves predictive performance over unimodal models, but the magnitude and consistency of improvement vary across learning outcomes and modality combinations. Among bimodal models, the combination of *Transcript+Audio* and *Transcript+Gallery* yields the highest gains, with F1-scores up to 0.69–0.70 on Social and Intrapersonal dimensions. These results suggest that the integration of linguistic content with vocal tone or visible social responsiveness provides a richer representation of communicative engagement and self-expression. The combinations *Audio+Screen* and *Audio+Gallery*, by contrast, yield smaller gains, indicating that perceptual cues without linguistic grounding capture limited cognitive information.

In three-modality settings, the inclusion of transcript features consistently enhances performance. Both *Transcript+Audio+Screen* and *Transcript+Audio+Gallery* exceed the best bimodal results, achieving up to 0.72 F1 in the Social and Intrapersonal categories. The latter combination is particularly effective, demonstrating that combining verbal, paralinguistic, and visible social cues captures not only cognitive processing but also interpersonal coordination—core attributes of collaborative learning in service-learning contexts. However, the gains diminish when adding a fourth modality: the full *T+A+S+G* model achieves the highest average F1 (0.73 in Social, 0.72 in Intrapersonal), but only modestly improves over the three-modality systems. This pattern reflects a saturation effect typical in multimodal learning analytics, where additional modalities yield diminishing returns once primary linguistic and behavioural channels are represented.

Interestingly, the superiority of the attention-based fusion strategy is not universal. While attention mechanisms produce the best overall results in most configurations, simple concatenation occasionally performs equally well or slightly better (e.g., in *Transcript+Audio* or *Screen+Gallery* combinations). This suggests that when training data are limited, the added parameterization of attention may introduce noise rather than consistent benefit, and straightforward integration can provide more stable generalisation. Nevertheless, attention-based fusion retains interpretive advan-

tages, as it enables dynamic weighting of modalities according to their contextual relevance for each prediction.

Overall, the findings confirm that multimodal fusion enhances the robustness and expressiveness of learning prediction models by integrating complementary behavioural signals across linguistic, paralinguistic, and visual domains. Language remains the dominant modality, but acoustic, screen-based, and visual cues together enrich the model's understanding of how cognitive, social, and affective processes jointly manifest in online service-learning environments. The results thus highlight the importance of balanced multimodal integration—where linguistic features anchor conceptual understanding, and visual and acoustic modalities provide behavioural context—to build reliable, scalable analytics for adaptive online education.

5.5.1 Limitations and contextual considerations

Several limitations of this study should be acknowledged. First, the participant sample comprises university students from Hong Kong serving as tutors to international service recipients from India, the Philippines, and South Africa. This cross-cultural composition introduces considerable variation in communicative norms, language proficiency, and interactional expectations. While this diversity reflects the authentic conditions of international service-learning, it also means that the behavioural patterns captured by the models may be shaped by culture-specific factors—such as norms around camera usage, verbal participation, and nonverbal expressiveness—that limit direct generalisability to other educational settings or cultural contexts.

Second, the theoretical mechanisms linking specific behavioural features to specific learning dimensions deserve further elaboration. The finding that linguistic features best predict intellectual learning gains is consistent with constructivist theories of learning, which posit that articulating ideas through discourse promotes cognitive restructuring and deeper understanding. Similarly, the association between camera-

usage features and social and civic learning outcomes aligns with theories of social presence, which suggest that visible participation fosters interpersonal trust, empathy, and a sense of shared purpose. However, the current predictive framework does not model these mediating processes explicitly; it establishes statistical associations rather than causal pathways. Future work could integrate mediation analysis or structural equation modelling to test whether specific behavioural indicators serve as proxies for theoretically motivated latent processes such as cognitive elaboration, social bonding, or reflective self-regulation.

Third, the reliance on self-reported learning gains, while standard in service-learning research, introduces the possibility that the models partially capture dispositional tendencies—such as general positivity or engagement orientation—rather than actual learning. Complementing self-reports with objective indicators such as project quality assessments, peer evaluations, or pre-post knowledge tests would provide a more robust basis for validating the behavioural predictors identified in this study.

5.6 Summary

This chapter presented a comprehensive multimodal analysis of synchronous online service-learning programs, aiming to predict tutors' self-reported learning gains from behavioural and communicative cues captured in videoconferencing data. Four modalities—transcript, audio, screen, and gallery view—were examined under both unimodal and multimodal configurations, offering complementary perspectives on linguistic, paralinguistic, instructional, and visual behaviour.

Unimodal analyses revealed that linguistic features were the most reliable predictors of learning gains, reflecting the central role of discourse coherence and metacognitive language in tutors' reflective learning processes. Audio features provided additional affective and expressive information, while screen content captured cognitive and in-

structional aspects related to knowledge presentation. The newly introduced gallery-view features extended these insights to nonverbal behaviour, showing that camera presence and body movement offer meaningful indicators of social engagement and civic awareness, albeit with weaker standalone predictive power.

Multimodal fusion further improved performance by integrating complementary behavioural signals. Models combining transcript, audio, and visual modalities consistently outperformed single-modality baselines, with the four-modality fusion achieving the highest accuracy on social and intrapersonal outcomes. The results underscore that learning in synchronous service-learning environments is jointly shaped by linguistic expression, vocal delivery, and visible participation. However, the marginal gains from additional modalities also suggest a limit to multimodal scalability when key behavioural dimensions are already well represented.

Overall, this study demonstrates that multimodal learning analytics can effectively capture the interplay between cognitive, social, and affective processes in online instruction. By linking automatically extracted behavioural features to validated measures of learning gain, the findings provide empirical foundations for developing adaptive feedback and analytics tools that support more responsive and reflective online teaching.

Chapter 6

Discussion

6.1 Evaluation and interpretation of experimental design

This thesis investigates how multimodal behavioural signals can be used to understand human interaction in group contexts. The three studies—backchannel detection, agreement estimation, and learning-gain prediction—were all designed to capture complementary levels of human communication: the occurrence of listener responses, the interpretation of their attitudinal meaning, and the relation between interactional behaviour and learning outcomes. Although each task adopts a distinct dataset and evaluation metric, they share the same underlying assumption: the observed behavioural patterns are reliable indicators of the participant's communicative or cognitive state. The validity of this assumption determines the interpretability of model performance.

In the backchannel detection and agreement estimation studies, the annotation process relies on observable nonverbal cues such as head motion, gaze, and facial expressions, combined with temporal alignment to the speaker's utterances. This approach

is effective for modelling spontaneous conversation, but it also inherits the ambiguity of human interpretation. For example, a head nod may indicate acknowledgment, agreement, or encouragement depending on the conversational context. By modelling the surrounding temporal context through hierarchical attention, the proposed temporal visual framework mitigates this ambiguity to some extent, yet annotation reliability remains bounded by subjective perception.

More broadly, the behavioural proxies used across all three studies may not map onto a single psychological construct. A head nod labelled as a backchannel could serve as a continuity signal, an expression of agreement, or a polite acknowledgment; similarly, sustained camera presence in the learning-gain study may reflect genuine social engagement, habitual behaviour, or compliance with institutional norms rather than deep cognitive involvement. These one-to-many mappings between observable behaviour and underlying communicative or cognitive intent represent a fundamental challenge in social signal processing. While the models in this thesis learn statistical regularities between behavioural features and annotated labels, the interpretive link between predicted outputs and psychological constructs should be treated with caution. Future work could address this limitation by incorporating richer annotation schemes that distinguish among communicative functions, or by triangulating model predictions with complementary measures such as retrospective self-reports or physiological indicators.

The learning-gain prediction task is grounded in self-reported outcomes, which provide a reasonable approximation of individual learning development but cannot fully capture objective achievement. While the self-report method is widely used in service-learning research, it may introduce response bias or cultural variation in self-assessment. To reduce such effects, this study focused on relative rather than absolute prediction accuracy. The use of behavioural rather than content-based features also helps ensure that the models reflect observable engagement patterns rather than linguistic self-presentation. Nevertheless, integrating external performance indicators, such

as project assessments or peer ratings, could further strengthen the validity of behavioural inference in future studies.

It should also be acknowledged that shared dispositional factors—such as confidence, engagement tendency, or self-enhancement bias—may simultaneously influence both behavioural expression and self-reported learning outcomes, introducing a potential common-method or latent-variable confound. For instance, a naturally outgoing tutor may speak more frequently, maintain camera presence, and also rate their own learning gains more favourably, without the behavioural features causally contributing to the learning outcome. Although the multimodal framework captures meaningful statistical associations, disentangling the causal pathways from behavioural indicators to learning outcomes would require designs that control for such dispositional variables, for example through longitudinal tracking, pre-post personality assessments, or instrumental variable approaches.

6.2 Comparative analysis of modelling approaches

Throughout the three studies, the choice of modelling strategy strongly influenced the ability to capture the nature of human behaviours. In backchannel detection, the proposed temporal visual attention framework demonstrated that multi-scale temporal modelling and channel-specific attention are essential for identifying brief yet meaningful visual signals. This design allows the model to attend to subtle motion variations that correspond to listener feedback, achieving higher accuracy than general-purpose vision models trained on large but unrelated datasets. The results suggest that human social cues require explicit temporal reasoning rather than broad-spectrum feature learning.

The agreement estimation task extended this design to multimodal integration. Combining visual descriptors (facial landmarks, head pose, and facial action units) with

prosodic features significantly improved prediction accuracy over unimodal baselines. The inclusion of speaker diarisation proved particularly useful: by isolating listener segments, it reduced noise introduced by the speaker's movement or speech and enabled the model to learn more consistent patterns of listener alignment. These findings highlight that context-aware preprocessing can be as important as model complexity in social signal tasks.

In learning-gain prediction, multimodal fusion offered a different set of insights. Linguistic and content-related features dominated the predictive performance, while visual indicators such as camera use and body posture provided complementary information about social presence. Compared with the first two studies, where attention-based fusion improved accuracy, simple concatenation performed competitively here. This can be attributed to data scale and noise: with limited samples per class, a lightweight integration strategy preserved complementary information without introducing additional parameters that might overfit. Together, these results demonstrate that the optimal modelling configuration depends on the granularity of the target construct and the reliability of available signals.

6.3 From prediction to behavioural explanation

The three studies in this thesis demonstrate that multimodal models can predict communicative events and learning outcomes with meaningful accuracy. However, predictive performance alone does not fully explain the behavioural processes that generate these outcomes. Understanding *why* certain features are predictive—and what they reveal about the social and cognitive mechanisms at work—is equally important for translating computational findings into actionable insights for education and human–computer interaction research.

In the backchannel detection study, the attention visualisations in Chapter 3 provide

a partial window into the model's reasoning: the Micro-CM and Macro-CM modules consistently assign higher weights to head movement and facial channels during confirmed backchannel events, aligning with the known communicative salience of nodding and smiling in listener feedback. These correspondences suggest that the model captures behaviourally meaningful patterns rather than spurious correlations, yet the framework does not explicitly model the conversational function of each detected backchannel—whether it serves as an acknowledgment, an encouragement, or an expression of agreement.

In the agreement estimation study, the fusion analyses reveal that energy-based acoustic features and fine-grained facial dynamics are the most informative predictors. These findings are consistent with communication theory, which holds that prosodic warmth and facial expressiveness signal positive affect and social alignment. Nevertheless, the regression model estimates a continuous agreement score without decomposing it into the constituent communicative acts—such as endorsement, compliance, or affective support—that may underlie the annotation. A richer explanatory account would require annotation schemes that capture these distinctions and models that can relate specific feature patterns to specific communicative functions.

In the learning-gain prediction study, the strongest predictors—utterance embeddings for intellectual outcomes, and camera-usage features for social and civic outcomes—point to plausible behavioural mechanisms. Linguistically rich and coherent discourse likely reflects deeper cognitive engagement and conceptual understanding, while sustained visual presence signals interpersonal commitment and collaborative readiness. Yet the current analysis does not establish whether these behaviours directly cause learning gains or merely co-occur with them. The theoretical linkage between conversational behaviour and multidimensional learning outcomes remains an area where future work could draw more explicitly on educational theories of experiential and service-learning to articulate the mediating processes.

Moreover, the naturalistic settings used across all three studies introduce ecological

variability that both strengthens and constrains interpretation. The authenticity of the data ensures that the models encounter realistic behavioural variation, but it also means that contextual factors—such as discussion topic, task structure, instructor facilitation style, and session design—may systematically shape the behavioural patterns observed. These factors are not controlled or modelled in the current framework, and their influence on the results should be considered when generalising the findings to other settings or populations.

6.4 Contextual and individual factors

The behavioural signals analysed in this thesis are not produced in isolation; they emerge within specific social, cultural, and interpersonal contexts that shape their form and interpretation. While the use of naturalistic data is a methodological strength, it also means that the findings are situated within particular contextual conditions that warrant explicit acknowledgment.

At the group level, interpersonal dynamics such as familiarity among participants, perceived social hierarchy, conversational dominance, and role distribution are likely to influence both backchannel frequency and the expression of agreement. In the MPI-IGroupInteraction dataset, for instance, some participants may have known each other prior to the recording, while others may have been strangers, potentially leading to different baseline levels of nonverbal responsiveness. Similarly, in the service-learning programmes, tutor teams with stronger pre-existing rapport may exhibit different interactional patterns than newly formed groups. These group-dynamic factors are not explicitly modelled or controlled in the current studies and may contribute to within-dataset variability that the models absorb as noise or as learned regularities that do not generalise to other group compositions.

At the individual level, personality traits and cognitive characteristics represent an-

other source of behavioural variability that the current framework does not account for. Factors such as introversion–extraversion, social anxiety, agreeableness, and neurodivergent characteristics (e.g., autism spectrum traits or ADHD-related tendencies) are known to influence backchannel frequency, facial expressiveness, gaze patterns, and participation willingness in group settings. For example, a highly introverted participant may produce fewer but more deliberate backchannels, while a participant with high agreeableness may nod frequently regardless of actual cognitive alignment. These individual differences could introduce systematic variability in the behavioural features that confounds the relationship between observed cues and the target constructs. Incorporating personality or trait-level covariates in future modelling efforts could help disentangle stable individual tendencies from context-dependent communicative signals.

The demographic and cultural composition of the participant samples also warrants consideration. The backchannel detection and agreement estimation studies draw on the MPIIGroupInteraction dataset (German university students) and the CCDb dataset (UK university participants), while the learning-gain study involves Hong Kong university tutors working with international service recipients from India, the Philippines, and South Africa. Research in cross-cultural pragmatics has established that backchannel frequency, agreement expression norms, and nonverbal display rules vary considerably across cultures and languages. Although the models achieve strong performance within the studied datasets, the extent to which these findings generalise across cultural and linguistic contexts remains an open question. This is particularly relevant for the agreement estimation task, where culturally specific norms around directness, politeness, and the social acceptability of disagreement may shape how agreement is expressed and perceived. Future studies could address this limitation by evaluating the proposed frameworks on multilingual or cross-cultural datasets and by incorporating culture-aware normalisation or adaptation mechanisms.

6.5 Extending multimodal features under the same framework

The experiments in this thesis primarily utilise visual, acoustic, and linguistic modalities. However, the same experimental framework could be extended to incorporate additional sources of behavioural information. Facial micro-expressions and gaze fixation patterns, already partially represented through OpenFace and ViTPose features, could be complemented by facial emotion rate or pupil dilation measures if camera quality allows. These cues are known to correlate with cognitive load and attention level and may refine agreement estimation and learning-gain prediction.

Similarly, body posture and hand gesture features could be further explored for group dynamics analysis. The current datasets focus on upper-body cues due to camera framing, yet gestural emphasis and torso orientation often reflect degrees of involvement. Future data collection could include full-body recordings to support such analysis. Moreover, incorporating physiological measures such as heart rate variability or galvanic skin response would provide an objective cross-check for inferred affective states, analogous to how gaze and mouse coordination were triangulated in earlier HCI studies.

6.6 Implications beyond the current scope

The contributions of this thesis extend beyond the specific models or datasets employed. From a methodological perspective, the work demonstrates that temporal encoding of fine-grained behaviours—through subtraction, differencing, and attention weighting—provides a general strategy for translating raw multimodal signals into interpretable social indicators. From a theoretical standpoint, the findings support the view that conversational responsiveness, agreement, and learning engagement are

governed by shared principles of synchrony and adaptation.

Practically, these findings suggest multiple avenues for application. In educational technology, automatic detection of responsive cues and agreement states could support real-time facilitation in online group learning, helping instructors monitor engagement without manual annotation. For instance, a dashboard that highlights moments of low backchannel activity or declining agreement could prompt instructors to adjust pacing, invite questions, or restructure group composition. In assessment contexts, the learning-gain prediction framework could complement traditional evaluation methods by providing continuous, behaviour-based indicators of engagement and participation, reducing reliance on end-of-course surveys alone.

In social robotics and virtual communication systems, adaptive models that interpret visual feedback and prosody could enhance naturalness and empathy in interaction. A conversational agent equipped with backchannel detection could generate appropriately timed listener responses, while agreement-aware systems could adapt their communication strategy based on the perceived level of user alignment. Beyond dyadic interaction, the multimodal frameworks developed here could inform the design of group collaboration tools that monitor and visualise participation balance, attentional coordination, and affective climate across team members.

The emphasis on interpretable temporal structures also offers a bridge between data-driven modelling and behavioural theory. By aligning computational attention maps with known communicative events, such as head nods or gaze shifts, the approach makes it possible to visualise and validate how models internalise behavioural meaning. This interpretability is essential for building trust in multimodal analytics and for integrating such systems into real educational and social applications.

6.7 Summary

This chapter has discussed the methodological, conceptual, and contextual implications of the three studies presented in this thesis. The evaluation of experimental design highlighted the construct validity challenges inherent in mapping observable behaviours to psychological constructs, and the potential for common-method confounds in self-reported learning outcomes. The comparative analysis of modelling approaches confirmed that the optimal configuration depends on the granularity and reliability of the target construct. The discussion of prediction and explanation underscored the importance of moving beyond performance metrics to articulate the behavioural mechanisms that underlie model predictions. The analysis of contextual and individual factors acknowledged that group dynamics, personality traits, and cultural norms represent unmodelled sources of variability that may influence both behavioural expression and model generalisability. Together, these reflections confirm that modelling short-term temporal patterns and selectively integrating complementary modalities are crucial for understanding human interaction in naturalistic settings. Context awareness, role differentiation, construct validity, and interpretability emerge as key design principles for future multimodal research. By bridging visual perception, acoustic modelling, and learning analytics, this work contributes to a broader framework for computational understanding of social interaction and its relation to human learning and collaboration.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis investigates how multimodal behavioural analysis can be used to understand human interaction and learning processes in group contexts. The three studies collectively examine a hierarchy of phenomena: (i) the occurrence of listener feedback in group conversation, (ii) the interpretation of agreement intensity expressed through backchannel behaviours, and (iii) the prediction of learning gains in online service-learning programmes. Together, these investigations provide a comprehensive view of how visual, acoustic, and linguistic signals reflect social responsiveness, attitudinal alignment, and learning engagement.

The first study establishes a temporal visual attention framework for detecting backchannel behaviours in multi-party conversation. By integrating micro-, uni-, and macro-level temporal modules, the framework captures both instantaneous and sustained visual dynamics, enabling accurate recognition of subtle listener actions such as nods and smiles. Evaluation on the MPIIGroupInteraction and CCDb datasets demonstrates that explicitly modelling temporal continuity and cross-channel relations yields substantial improvement over baseline and general-purpose vision models. This find-

ing confirms that fine-grained temporal reasoning is indispensable for decoding spontaneous nonverbal cues in natural group interactions.

Building on this foundation, the second study extends the focus from behaviour recognition to attitudinal interpretation. A multimodal regression framework is developed to estimate agreement intensity by combining visual and acoustic features. The results show that visual cues—particularly head orientation, facial actions, and gaze patterns—are reliable indicators of agreement, while energy- and spectral-based audio descriptors provide complementary information about prosodic modulation. The incorporation of speaker diarisation improves consistency by isolating listener-specific behaviours from speaker activity. These findings indicate that agreement is best inferred from synchronised, short-term patterns across modalities rather than from isolated signals, highlighting the role of multimodal convergence in modelling subtle affective states.

The third study connects micro-level behavioural modelling with macro-level learning analytics. Using multimodal data from synchronous service-learning sessions, it predicts students' intellectual, social, civic, and intrapersonal learning gains from conversational features. The results reveal that linguistic indicators derived from transcripts are the strongest predictors of intellectual outcomes, while camera-usage and pose-difference features from gallery-view videos contribute significantly to social and civic learning. Screen-content embeddings further capture cognitive aspects linked to topic depth and task complexity. These results suggest that learning effectiveness in collaborative online settings can be inferred from behavioural markers of verbal engagement and visual participation.

These studies demonstrate a coherent methodological progression—from visual temporal analysis, to multimodal attitudinal inference, to outcome-level prediction—anchored in the same principle that short-horizon temporal modelling and selective multimodal integration are essential for understanding human interaction. The thesis contributes both methodological advances, such as motion-aware temporal attention and role-

aware preprocessing, and conceptual insights into how responsiveness, agreement, and engagement are interlinked across communicative and educational domains.

7.2 Future Work

Although the present studies confirm the feasibility of multimodal modelling for social and learning behaviours, several directions remain open for future exploration.

(1) Expansion of datasets and domains. The current datasets, while diverse in modality and context, are limited in scale and demographic coverage. Enlarging the participant pool and extending to additional cultural or linguistic backgrounds would allow more generalisable conclusions about behavioural variability. For educational contexts, expanding the corpus to multiple courses or disciplines could test the robustness of the learning-gain predictors across pedagogical settings.

(2) Integration of additional modalities. Future work may incorporate complementary physiological or behavioural signals such as eye-blink rate, pupil dilation, or heart-rate variability. These signals can provide objective measures of cognitive load and emotional arousal that enrich interpretation of backchannel and agreement behaviours. For learning analytics, combining such measures with video and transcript features could help disentangle affective engagement from task difficulty.

(3) Adaptive and data-efficient modelling. While the proposed frameworks perform well on moderate-sized datasets, large-scale deployment requires models that adapt efficiently to new domains. Few-shot or self-supervised learning methods could be explored to transfer representations from one interaction context to another. In educational applications, adaptive models could learn personalised baselines for each learner or group, similar to the transition from user-independent to user-dependent settings in HCI studies.

(4) Modelling individual differences and contextual factors. As discussed in

Chapter 6, the current studies do not explicitly account for individual difference variables such as personality traits (e.g., introversion–extraversion, agreeableness), social anxiety, or neurodivergent characteristics, all of which are known to influence nonverbal expressiveness and participation patterns. Incorporating such covariates—either as input features, as conditioning variables, or through personalised normalisation layers—could improve model robustness and help disentangle stable individual tendencies from context-dependent communicative signals. Similarly, modelling group-level factors such as familiarity, role hierarchy, and conversational dominance could enhance the interpretability of backchannel and agreement behaviours in multi-party settings.

(5) Cross-cultural and cross-linguistic evaluation. The datasets employed in this thesis are drawn from Western European (German and UK) and East Asian (Hong Kong) contexts, limiting the cultural scope of the findings. Backchannel norms, agreement expression, and nonverbal display rules vary substantially across cultures and languages. Evaluating the proposed frameworks on multilingual or cross-cultural datasets would test the generalisability of the learned representations and identify culture-specific behavioural patterns that require adaptation. Developing culture-aware models—for example, through domain adaptation or culture-conditioned normalisation—represents a promising direction for building more inclusive multimodal interaction systems.

(6) Application to broader interaction domains. While this thesis focuses on group conversation and educational settings, the multimodal analysis frameworks developed here have potential relevance to other domains where group interaction plays a central role. In team sports, for instance, nonverbal coordination and backchannel-like feedback among players may relate to team cohesion and performance. In crowd behaviour analysis, detecting patterns of collective agreement or attentional synchrony could support safety monitoring and event management. In clinical and therapeutic settings, modelling patient–therapist interactional dynamics could inform assessments

of rapport, engagement, and treatment responsiveness. Exploring these applications would not only test the generalisability of the proposed methods but also broaden the impact of multimodal interaction research beyond its current boundaries.

In conclusion, this thesis demonstrates that multimodal and temporally aware computational modelling can provide deep insights into human interaction and learning. By bridging the analysis of micro-behavioural cues and macro-level outcomes, it lays a foundation for the development of socially aware and educationally meaningful intelligent systems. Future extensions that expand the scope of data, modalities, and interpretive granularity will further enhance our understanding of how people communicate, align, and learn together.

Chapter 8

Other Contributions

Beyond the core studies presented in this thesis, several additional research efforts have extended its methodological principles into related domains. These works share a common objective of advancing multimodal understanding and adaptive representation learning, spanning areas such as social signal processing, trustworthy communication, educational data mining, and scientific representation learning. Collectively, they demonstrate the versatility of temporal modelling and multimodal integration across diverse applications.

Multimodal Behaviour Analysis and Social Signal Processing

Engagement estimation in human interaction. A lightweight multimodal framework was developed to estimate engagement levels in dyadic conversation by combining body motion descriptors with short-term audio features. The system captured temporal synchrony between interlocutors and achieved substantial improvement over existing baselines. Analyses showed that longer observation windows and the inclusion of partner features enhanced model stability, reinforcing the importance of contextual and interpersonal information for engagement estimation.

Backchannel agreement estimation with individual standardisation. The CMIS-Net architecture introduced a cascaded multi-scale individual standardisation mechanism to separate person-specific and invariant behavioural patterns. This approach reduced inter-personal variability and mitigated data imbalance in agreement-intensity prediction. Results demonstrated improved generalisation across speakers and backchannel types, particularly for auditory signals, highlighting the value of adaptive normalisation in modelling heterogeneous social behaviours.

Trustworthy AI for Communication Security

Privacy-preserving phone-scam detection with modular sanitisation. A modular adaptive sanitisation framework, MASK, was proposed for privacy-preserving phone-scam detection using large language models. The system supports configurable sanitisation strategies—from keyword masking to neural redaction—allowing flexible trade-offs between privacy protection and detection accuracy. The framework contributes a general architecture and training paradigm for balancing transparency, reliability, and data protection in real-world conversational AI applications.

LLM-based scam detection framework. A conceptual analysis of LLM-based scam-detection pipelines was conducted to identify open challenges in robustness, multilingual generalisation, and privacy. This position work situates the MASK framework within a broader research landscape and outlines future directions for developing trustworthy, secure, and adaptable conversational intelligence systems.

Educational Data Mining and Recommender Systems

Multimodal-enhanced collaborative filtering for e-learning A multimodal collaborative-filtering model was developed to support personalised learning-resource recommendation. The framework integrates learners' behavioural histories, seman-

tic representations of learning materials, and temporal activity dynamics to model evolving preferences. A learner–resource importance weighting and recurrent temporal component improve both recommendation precision and interpretability, contributing to adaptive learning analytics in online education.

Molecular Representation Learning

Molecular property prediction. The Ring2Vec representation-learning method was introduced to model molecular substructures as distributional embeddings. By encoding atomic composition and relational context in a unified vector space, the approach enables efficient and accurate prediction of key molecular properties in organic materials. This work extends linguistic-inspired embedding techniques to scientific data, demonstrating the generalisability of representation-learning principles beyond human communication.

Across these studies—spanning human interaction analysis, secure communication, educational recommendation, and molecular modelling—the unifying principle lies in representation design that aligns structure, context, and meaning. Each contribution operationalises the core methodological ideas of this thesis: temporal modelling of sequential dynamics, multimodal fusion of complementary cues, and adaptive standardisation for individual or contextual variability. Collectively, these works demonstrate the broader applicability and impact of the research framework developed in this dissertation.

References

[1] Halim Acosta, Seung Lee, Bradford Mott, Haesol Bae, Krista Glazewski, Cindy Hmelo-Silver, and James Lester. Multimodal learning analytics for predicting student collaboration satisfaction in collaborative game-based learning. 2024.

[2] Ahmed Amer, Chirag Bhuvaneshwara, Gowtham K Addluri, Mohammed M Shaik, Vedant Bonde, and Philipp Müller. Backchannel detection and agreement estimation from video with transformer networks. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

[3] Michael Argyle. *Bodily communication*. Routledge, 2013.

[4] Michael Argyle and Mark Cook. Gaze and mutual gaze. 1976.

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[6] Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 277–282, 2013.

[7] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[9] Daniel R Bailey, Norah Almusharraf, and Asma Almusharraf. Video conferencing in the e-learning context: explaining learning outcome with the technology acceptance model. *Education and Information Technologies*, 27(6):7679–7698, 2022.

[10] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[11] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[12] David Baron, Andrew Bestbier, Jennifer M Case, and Brandon I Collier-Reed. Investigating the effects of a backchannel on university classroom interactions: A mixed-method case study. *Computers & Education*, 94:61–76, 2016.

[13] German Barquero, Johnny Núnez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 139–178. PMLR, 2022.

[14] Janet B Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941, 2000.

[15] Eric S Belt and Patrick R Lowenthal. Synchronous video-based communication and online learning: An exploration of instructors' perceptions and experiences. *Education and Information Technologies*, 28(5):4941–4964, 2023.

[16] Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. Moving as a leader: Detecting emergent leadership in small groups using body pose. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1425–1433, 2017.

[17] Maneesh Bilalpur, Mert Inan, Dorsa Zeinali, Jeffrey F Cohn, and Malihe Alikhani. Learning to generate context-sensitive backchannel smiles for embodied ai agents with applications in mental health dialogues. *arXiv preprint arXiv:2402.08837*, 2024.

[18] Kübra Bodur, Mitja Nikolaus, Laurent Prévot, and Abdellah Fourtassi. Using video calls to study children's conversational development: The case of backchannel signaling. *Frontiers in Computer Science*, 5:1088752, 2023.

[19] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–9. IEEE, 2009.

[20] Konstantinos Bousmalis, Louis-Philippe Morency, and Maja Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 746–752. IEEE, 2011.

[21] Lawrence J Brunner. Smiles can be back channels. *Journal of personality and social psychology*, 37(5):728, 1979.

[22] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm:

Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[23] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[24] Doug Clow. An overview of learning analytics. *Teaching in Higher Education*, 18(6):683–695, 2013.

[25] Shuki Cohen. A computerized scale for monitoring levels of agreement during a conversation. 2003.

[26] Ben Daniel. B ig d ata and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, 46(5):904–920, 2015.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[28] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[29] Starkey Duncan and Donald W Fiske. *Face-to-face Interaction: Research, Methods, and Theory*. Routledge, 2015.

[30] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[31] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[32] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[33] Chris D Frith and Uta Frith. Mechanisms of social cognition. *Annual review of psychology*, 63(1):287–313, 2012.

[34] Eugene Yujun Fu, Grace Ngai, Hong Va Leong, Stephen CF Chan, and Daniel TL Shek. Using attention-based neural networks for predicting student learning outcomes in service-learning. *Education and Information Technologies*, pages 1–27, 2023.

[35] Maria Fusaro, Paul L Harris, and Barbara A Pan. Head nodding and head shaking gestures in children's early communication. *First Language*, 32(4):439–458, 2012.

[36] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and vision computing*, 27(12):1775–1787, 2009.

[37] Sarah Gillet, Marynel Vázquez, Christopher Peters, Fangkai Yang, and Iolanda Leite. Multiparty interaction between humans and socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, pages 113–154. 2022.

[38] Emer Gilmartin, Benjamin R Cowan, Carl Vogel, and Nick Campbell. Explorations in multiparty casual social talk and its relevance for social human machine dialogue. *Journal on Multimodal User Interfaces*, 12:297–308, 2018.

[39] Charles Goodwin. Between and within: Alternative sequential treatments of continuers and assessments. *Human studies*, 9(2-3):205–217, 1986.

[40] Maaike Grammens, Michiel Voet, Ruben Vanderlinde, Lieselot Declercq, and Bram De Wever. A systematic review of teacher roles and competences for

teaching synchronously online through videoconferencing technology. *Educational Research Review*, 37:100461, 2022.

[41] Agustín Gravano and Julia Hirschberg. Backchannel-inviting cues in task-oriented dialogue. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[42] Bettina Heinz. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics*, 35(7):1113–1142, 2003.

[43] Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*, pages 34–36, 2003.

[44] Hung-Hsuan Huang, Seiya Kimura, Kazuhiro Kuwabara, and Toyoaki Nishida. Generation of head movements of a robot using multimodal features of peer participants in group discussion conversation. *Multimodal Technologies and Interaction*, 4(2):15, 2020.

[45] Vidit Jain, Maitree Leekha, Rajiv Ratn Shah, and Jainendra Shukla. Exploring semi-supervised learning for predicting listener backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[46] Dinesh Babu Jayagopi, Hayley Hung, Chuohao Yeo, and Daniel Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.

[47] Zoltan Katai and David Iclanzan. Impact of instructor on-slide presence in synchronous e-learning. *Education and Information Technologies*, 28(3):3089–3115, 2023.

[48] Mary Ritchie Key. *The relationship of verbal and nonverbal communication*. Number 25. Walter de Gruyter, 1980.

[49] Olga M Klibanov, Christian Dolder, Kevin Anderson, Heather A Kehr, and J Andrew Woods. Impact of distance education via interactive videoconferencing on students' course performance and satisfaction. *Advances in physiology education*, 42(1):21–25, 2018.

[50] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Thomson Wadsworth, 1972.

[51] Masato Kogure. Nodding and smiling in silence during the loop sequence of backchannels in japanese conversation. *Journal of Pragmatics*, 39(7):1275–1289, 2007.

[52] George Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.

[53] Han Z Li. Backchannel responses as misleading feedback in intercultural discourse. *Journal of intercultural communication research*, 35(2):99–116, 2006.

[54] Jun Li, Xianglong Liu, Wenxuan Zhang, Mingyuan Zhang, Jingkuan Song, and Nicu Sebe. Spatio-temporal attention networks for action recognition and detection. *IEEE Transactions on Multimedia*, 22(11):2990–3001, 2020.

[55] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[56] Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, Liangyin Chen, and Yanru Chen. Lr-asd: Lightweight and robust network for active speaker detection. *International Journal of Computer Vision*, 133(7):4749–4769, 2025.

[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[58] Yingbo Ma, Gloria Ashiya Katuka, Mehmet Celepkolu, and Kristy Elizabeth Boyer. Investigating multimodal predictors of peer satisfaction for collaborative coding in middle school. *International educational data mining society*, 2022.

[59] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 638–647, 2022.

[60] Katerina Mangaroska, Kshitij Sharma, Dragan Gaševic, and Michalis Giannakos. Multimodal learning analytics to inform learning design: Lessons learned from computing education. *Journal of Learning Analytics*, 7(3):79–97, 2020.

[61] Elisabetta Mazzullo, Okan Bulut, Tarid Wongvorachan, and Bin Tan. Learning analytics in the era of large language models. *Analytics*, 2(4):877–898, 2023.

[62] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[63] Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous agents and multi-agent systems*, 20:70–84, 2010.

[64] Desmond Morris. Man watching. *A field guide to human behaviour*, 1977.

[65] Su Mu, Meng Cui, and Xiaodi Huang. Multimodal data fusion in learning analytics: A systematic review. *Sensors*, 20(23):6856, 2020.

[66] Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. Multimediate'22: Backchannel detection and agreement estimation in group interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7109–7114, 2022.

[67] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*, pages 153–164, 2018.

[68] Grace Ngai, Stephen CF Chan, and Kam-por Kwan. Challenge, meaning and preparation: Critical success factors influencing student learning outcomes from service-learning. *Journal of Higher Education Outreach and Engagement*, 22(4):55–80, 2018.

[69] Naomi Nota, James P Trujillo, and Judith Holler. Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11(8):1017, 2021.

[70] Xavier Ochoa, Charles Lang, George Siemens, Alyssa Wise, Dragan Gasevic, and Agathe Merceron. Multimodal learning analytics-rationale, process, examples, and direction. *The handbook of learning analytics*, pages 54–65, 2022.

[71] Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. Oh, jeez! or uh-hah? a listener-aware backchannel predictor on asr transcriptions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8064–8068. IEEE, 2020.

[72] Sandra Ottl, Shahin Amiriparian, Maurice Gerczuk, Vincent Karas, and Björn Schuller. Group-level speech emotion recognition utilising deep spectrum features. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 821–826, 2020.

[73] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas Huang. Human computing and machine understanding of human behavior: A survey. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 239–248, 2006.

[74] Zhongling Pi, Li Zhang, Xin Zhao, and Xiying Li. Peers turning on cameras promotes learning in video conferencing. *Computers & Education*, 212:104986, 2024.

[75] Vojtěch Pipek. *On backchannels in English conversation*. PhD thesis, Masarykova univerzita, Pedagogická fakulta, 2007.

[76] Isabella Poggi, Francesca D'Errico, Laura Vincze, et al. Types of nods. the polysemy of a social signal. In *LREC*, 2010.

[77] Isabella Poggi, Francesca D'Errico, and Laura Vincze. Agreement and its multimodal communication in debates: A qualitative analysis. *Cognitive Computation*, 3:466–479, 2011.

[78] Yongfeng Qi, Liqiang Zhuang, Huili Chen, Xiang Han, and Anye Liang. Evaluation of students' learning engagement in online classes based on multimodal vision perspective. *Electronics*, 13(1):149, 2023.

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[80] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[81] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[82] Virginia P Richmond. Nonverbal behavior in interpersonal relations. *(No Title)*, page 366, 2008.

[83] Garima Sharma, Kalin Stefanov, Abhinav Dhall, and Jianfei Cai. Graph-based group modelling for backchannel detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7190–7194, 2022.

[84] Apeksha Shewalkar, Deepika Nyavanandi, and Simone A Ludwig. Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru. *Journal of Artificial Intelligence and Soft Computing Research*, 9(4):235–245, 2019.

[85] Elizabeth Shriberg. Spontaneous speech: how people really talk and why engineers should care. In *INTERSPEECH*, pages 1781–1784. Citeseer, 2005.

[86] Nikhita Singh, Jin Joo Lee, Ishaan Grover, and Cynthia Breazeal. P2pstory: Dataset of children as storytellers and listeners in peer-to-peer interactions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

[87] Angela EB Stewart, Zachary A Keirn, and Sidney K D'Mello. Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 21–30, 2018.

[88] Tanya Stivers. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on language and social interaction*, 41(1):31–57, 2008.

[89] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*, 2020.

[90] Jackson Tolins and Jean E Fox Tree. Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164, 2014.

[91] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[92] Khiet P Truong, Ronald Poppe, and Dirk Heylen. A rule-based backchannel prediction model using pitch and pause information. In *Eleventh Annual Conference of the International Speech Communication Association*. Citeseer, 2010.

[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[94] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[95] Kangzhong Wang, MK Michael Cheung, Youqian Zhang, Chunxi Yang, Peter Q Chen, Eugene Yujun Fu, and Grace Ngai. Unveiling subtle cues: Backchannel detection using temporal multimodal attention networks. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9586–9590, 2023.

[96] Kangzhong Wang, Eugene Yujun Fu, Grace Ngai, and Hong Va Leong. Identifying key learning factors in service-learning programs using machine learning. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1312–1317. IEEE, 2022.

[97] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.

[98] Wen Wang, Sibel Yaman, Kristin Precoda, and Colleen Richey. Automatic identification of speaker role and agreement/disagreement in broadcast conversation. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5556–5559. IEEE, 2011.

[99] Sheida White. Backchannels across cultures: A study of americans and japanese1. *Language in society*, 18(1):59–76, 1989.

[100] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559, 2006.

[101] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[102] Victor H Yngve. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578, 1970.

[103] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.