TGN-SMM: A Temporal Graph-Based Shared Mental Model for Modelling Evolving Team Dynamics in Multi-Party Meetings

1st Moinam Chatterjee
Cognitive Systems Group
Institute for Information and Communication Technology,
Otto von Guericke University Magdeburg
Magdeburg, Germany
moinam.chatterjee@ovgu.de

2nd Behnam Ensan
Cognitive Systems Group
Institute for Information and Communication Technology,
Otto von Guericke University Magdeburg
Magdeburg, Germany
behnam.ensan@ovgu.de

3rd Ayoub Al-Hamadi
Neuro-Information Technology
Institute for Information and Communication Technology,
Otto von Guericke University Magdeburg
Magdeburg, Germany
ayoub.al-hamadi@ovgu.de

4th Andreas Wendemuth
Cognitive Systems Group
Institute for Information and Communication Technology,
Otto von Guericke University Magdeburg
Magdeburg, Germany
andreas.wendemuth@ovgu.de

***Abstract*—Understanding the collaborative behaviour between team members and the formation of shared mental models in multi-party meetings is essential to achieving shared teaming goals and effective Human-Agent Teaming. We introduce TGN-SMM, a Shared Mental Model framework based on a Temporal Graph Neural Network, designed to model the team's cognitive states in meetings as an evolving speaker-interaction graph. Each node here represents a team member, characterised by the context of their utterances, their team role, and the intention of their speech. As the meeting unfolds, pairwise temporal events are passed to update the speakers' cognitive states, and the interpersonal alignment dynamic is directly inferred from edge representations between speaker pairs. The interactions in the meeting are treated as pairwise temporal events, with the edge representation relying on local reasoning rather than a global graph representation. This allows the framework to capture how the cognitive states and interpersonal alignment dynamic patterns emerge and evolve. Temporal adjacency pairs in meetings are represented as a list of dyadic conversations among team members in the AMI Meeting Corpus, which is used to conduct experiments in which the TGN-SMM achieves the highest macro-averaged F1 score, with particular improvement on minority alignment classes on its static and temporal SMM baselines, illustrating the benefits of incorporating temporal graph reasoning into models of team alignment.**

***Index Terms*—Shared Mental Model, Temporal Shared Mental Model, Graph-Temporal Reasoning, Human-Agent Teaming, Multi-Party Conversation Modelling, Interpersonal Alignment Dynamic, Team Dynamics, AMI Meeting Corpus**

I. Introduction

In today's world, we see the widespread presence of versatile Artificial Intelligence (AI) agents that are capable of supporting active decision-making across a wide range of domains [1]. One such domain is Human-Agent Teaming Systems (HATs) [2], where AI agents are given the role of a teammate and are tasked to understand the conversational teaming context and dynamically adapt to evolving team dynamics to help facilitate effective collaboration and support active decision-making towards reaching shared team goals [3]. For an agent to assist proactively, it must not only be able to map the context of human team conversations but also their underlying cognitive states. To facilitate effective collaboration in human teams and that of human-agent teams, Shared Mental Models (SMMs) are introduced [4]. It is a foundational concept that plays a critical role in substantiating team cognition, referring to the common understanding among team members regarding tasks, roles, and interpersonal dynamics [5]. These cognitive representations shape the team members' dynamic cognitive states, which can shift over time and across conversations, thus capturing team dynamics.

To observe how an SMM unfolds in a real-time team collaboration scenario, multi-party meetings can be used, as they provide a rich team-dynamics environment involving teams of humans working towards shared goals, as captured in the AMI Meeting Corpus [6]. In these multi-party meetings, participants build and update their shared understanding through active collaboration, highlighting interpersonal team dynamics that can help them converge on decisions. The interpersonal team dynamic in the AMI Corpus captures moments of support, objection, and uncertainty among team members that signal whether the team is converging or diverging and showcase evolving alignments. This makes such multi-party meetings an ideal testbed for observing the real-time evolution of shared cognition, essential for an AI agent to function as a competent teammate, interpreting and enabling proactive assistance and effective collaboration [7].

Although Shared Mental Models (SMMs) hold a central role in effective teamwork, most computational models remain static, focusing on end-state agreement rather than showing temporal processes that build shared understanding over time [8]. Studies show that computational frameworks for team cognition tend to infer SMMs from static snapshots of communication or global behavioural summaries [9], overlooking the temporal shifts in interpersonal dynamics. These static representations fail to capture the fine-grained temporal evolution of team dynamics, rendering them insufficient for AI agent teammates to understand and learn the evolving cognitive states that characterise real collaborative interactions.

To address these limitations, we propose TGN-SMM, a Temporal Graph Neural framework designed to infer shared cognitive states and interpersonal dynamics in multi-party meetings. This framework is based on our previously proposed conceptual temporal shared mental model framework [10]. Our framework models meetings as an evolving speaker-interaction graph, where nodes represent a team member, characterised by the context of their utterances, team role, and the intention of their speech. TGN-SMM infers dyadic interpersonal alignments from the edge representation, allowing fine-grained modelling of interpersonal relations without relying on the global graph representation. We train and evaluate the framework on the meeting sessions of the AMI Meeting Corpus [6] by treating the dataset's annotated dyadic adjacency pairs as temporal events in the meetings. We evaluate how well the framework captures the temporal evolution of the team's cognitive states to infer the interpersonal alignments between team members.

II. Related Work

Research on conceptual frameworks such as Shared Mental Models (SMMs) provides the foundation for understanding how cognitive and shared understanding are maintained. Classical SMMs emphasise that effective teams rely on aligned cognitive representations of tasks, roles, and strategies [8], [11]. However, these models are typically assessed using static measures and structured questionnaires [12], which capture only the end-state convergence rather than the dynamic evolution of shared understanding during interaction. Formulations such as the Interactive Team Cognition [9] argue that team cognition is a direct product of communication among team members rather than of stored mental states. Yet, computational methods for capturing these dynamics remain limited.

Prior work on the addition of temporal team mental models argues that teams develop a shared understanding of task sequences and goals over time, thereby extending the SMM theory of ”what” and ”how” to the time domain and incorporating ”when” [13]. Empirical studies show that teams with stronger temporal SMMs coordinate more efficiently under time pressure and adapt more effectively to disruptions [14]. However, these models are still measured at discrete time points through surveys rather than through continuous analysis of interaction data. As a result, temporal SMM research highlights the importance of timing and sequencing but cannot provide a computational mechanism to capture how temporal expectations evolve within a conversation.

Research has also been conducted to understand how team learning behaviours interact with SMMs over time. Temporal common ground highlights how shared mental models maintain relationships among different team learning dynamics and how these relationships affect performance gains across repeated interactions [15]. The studies here highlight that the SMMs not only evolve but also shape the team's learning patterns across temporal interaction data. Yet the work relies on session-level measurements that show how SMMs model tasks before and after a timed session, capturing session-to-session evolution. This raises the question of how SMMs map fluctuating relationships within the team at the utterance level, how their interpersonal convergence signals contribute to SMM formation, and how these dynamics can be automatically inferred from temporal data.

There have been efforts to develop computational models that capture shared knowledge dynamics or team cognition using latent or network-based representations. Shared semantic spaces and concept maps are examples of computational models used to approximate team knowledge similarity [16], [17]. Prior work analyses temporal patterns in team interaction sequences, showing how communication dynamics reflect the formation of team cognition and coordination [18]. However, they rely on manually elicited concepts or temporal analysis windows, which may limit their ability to capture fine-grained transitions in the team's cognitive states during natural conversation.

Parallel research has examined how linguistic, acoustic, and visual cues reflect group processes in multimodal meetings. Studies have examined decision-making dynamics [19], dialogue structure [20], dominance and influence [21], and social dynamics such as agreement in conversation [22]. In addition, methods for recognising and interpreting multiparty meetings using multimodal signals have been developed based on the work on the AMI corpus [23]. These works demonstrate the richness of multi-party meeting data but often focus on specific aspects of dynamics, such as decision-making, dominance, or agreement, rather than explicitly modelling team-level alignment and the team's shared cognitive states.

In early dialogue systems, there was a major reliance on hierarchical recurrent architectures to capture multi-turn dependencies [24], whereas later Transformer-based models demonstrated strong performance in modelling long-range dependencies [25]. However, many of these approaches model conversational data as linear sequences and fail to explicitly capture the relational structure inherent in multi-party interaction data. To incorporate speaker-specific characteristics, persona-based models were introduced [26]. As a result, such models are limited in their ability to capture how the interpersonal alignments between team members evolve during collaborative tasks.

Prior work shows that adding a graph-based approach addresses some limitations by modelling conversations as structured interaction graphs. DialogueGCN [27] showed that convolutional graph networks can capture the speaker relationships and contextual dependencies for emotion recognition in conversations. Other works have utilised graph structures for discourse parsing [28] and for modelling multi-party dialogue as graphs, where interactions among members are represented as graph edges [29]. However, many of these models operate on static graphs having a fixed relational structure and do not explicitly capture how interactions evolve. To address this, temporal modelling frameworks are introduced. Temporal Graph Network (TGN) [30] is an example that extends graph neural networks (GNNs) to handle time-evolving graphs, in which event-based message passing and memory mechanisms enable node and edge representations to evolve as new interactions are encountered. Temporal models are also used for dynamic link prediction in evolving social networks [31], temporal recommendation systems [32], and temporal knowledge graph reasoning and event prediction [33]. However, in the application of modelling multi-party meetings, modelling shared cognitive states remains limited.

To map these fine-grained team dynamics, a multi-modal, multi-party meeting dataset is crucial. Multi-party meeting corpora such as the AMI Meeting Corpus [6] have allowed analysis on conversational dynamics, including role recognition and meeting dynamic understanding [23], topic segmentation using probabilistic and Bayesian approaches [34], decision detection and identification of decision points in meetings [19], and multimodal interaction analysis [6]. However, little work has focused on modelling the shared cognitive states and interpersonal team dynamics as evolving constructs within these meetings. As stated before, existing computational approaches tend to classify global meeting convergence, leaving a gap for methods such as temporal SMMs that can capture evolving temporal and relational dynamics.

Overall, the approaches from prior works provide valuable insights into team dynamics, such as conversational patterns and global team convergence outcomes; they do not jointly model the evolution of shared cognitive states and interpersonal alignment at a fine-grained, utterance-level event. In particular, existing methods rely on static or session-level assessments (e.g., questionnaires) that model conversational behaviour as a linear sequence rather than its temporal evolution. As a result, there is limited ability to capture how alignment emerges, shifts, and propagates among team members in natural conversation. This highlights a gap for a unified temporal-relational framework that can model team dynamics while simultaneously tracking the evolution of shared cognitive mental representations in multi-party meetings. This gap motivates the development of TGN-SMM.

III. Methodology

In this section, we explain our proposed Temporal Graph Neural Network Shared Mental Model (TGN-SMM), shown in Figure 1, designed to track the evolving cognitive states of team members during multi-party meetings as interactions unfold. Each meeting is organised as a list of temporal events that reflect conversations over time among team members. The model sees one temporal event between two team members in the meeting at a given time, used to maintain the Speaker-specific hidden state that evolves through:

- **Speaker embeddings (role, intention, text):** Each speaker is represented by a concatenated vector encoding their team role, intention class (dialogue act), and utterance semantics.
- **Alignment valence between team members:** Directional relational alignment is computed using pairwise feature interactions and a gated MLP to capture Speaker-to-Speaker similarity.
- **Dyadic Message passing:** Speaker cognitive states and alignment valences are combined through an MLP to generate interaction-specific messages.
- **Temporal update via a GRU:** Each team member's hidden state is updated using a GRU that integrates the incoming message with their previous state.

The resulting cognitive states of the interacting team members and their edge representation are used to predict the alignment class of each event in the meeting. The model architecture of TGN-SMM is shown in Figure 1, comprising the components explained in detail below.

A. Speaker Embeddings

For each speaker the model encounters in the meeting, a corresponding speaker cognitive state is constructed. As the model encounters a new speaker, a cognitive state is created with a default representation (new node). Thus, making the creation of cognitive states dynamic and the evolving graph independent of the number of team members in the meeting.

a) Role Embedding.: Functional team roles (e.g., *Project Manager*, *Marketing Expert*, etc.) encode each team member's functional responsibilities in the meeting. To map the hierarchical and functional structure in the meeting, we map each discrete role label to a dense vector:

$$\mathbf{role}_i = \mathrm{RoleEmb}(role_i) \in \mathbb{R}^{d_r}, \tag{1}$$

where d_r is the dimensionality of the role embedding

This allows the model to learn role-specific behavioural patterns and expectations.

b) Intention Embedding.: The Intention labels describe the dialogue act of the utterance made by the team member, providing context for its meaning (e.g., *Inform*, *Suggest*, etc.). These labels provide a high-level abstraction of the speaker's communicative goal. We embed each intention class into a continuous space:

$$\mathbf{intent}_i = \mathrm{IntentEmb}(intent_i) \in \mathbb{R}^{d_{da}}, \tag{2}$$

where d_{da} is the dimensionality of the role embedding

This enables the model to capture relationships between different communicative intents and how they influence alignment.

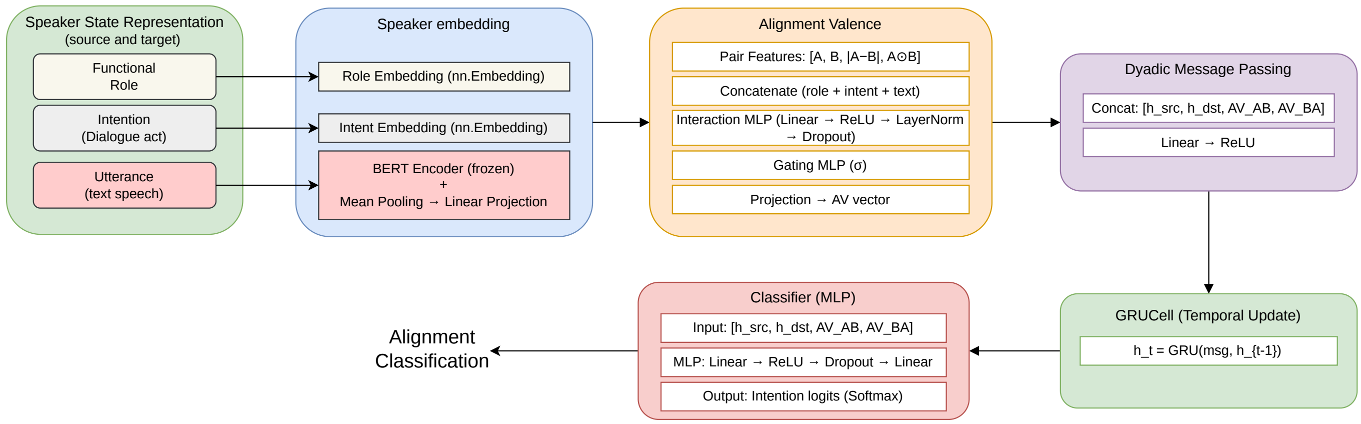

Fig. 1. TGN-SMM: Model Architecture

c) Text Embedding.: The semantic meaning of the utterance is encoded with the help of a frozen BERT [35], which is then followed by a projection and layer normalisation:

$$\mathbf{utter}_i = \mathrm{LN}(W_{\text{text}} \cdot \mathrm{BERT}(x_i)) \in \mathbb{R}^{d_u}. \tag{3}$$

where d_u is the dimensionality of the text embedding

This representation captures lexical, syntactic, and contextual information from the utterance.

d) Final Speaker Representation.: These three features represent the cognitive state of a team member in a conversation at time t in the Meeting M. They are concatenated to form the complete cognitive state of a team member at time t:

$$\mathbf{e}_i = [\mathbf{role}_i \,\|\, \mathbf{intent}_i \,\|\, \mathbf{utter}_i] \in \mathbb{R}^{d_e}, \tag{4}$$

where $d_e = d_r + d_{da} + d_u$.

This multimodal vector provides a rich description of each member's conversational identity and serves as the basis for computing alignment valence.

B. Alignment Valence

The alignment valance component maps the interpersonal dynamics between two team members at each conversational turn. The nature of the conversation is quantified by the degree to which two members align during an interaction. This module computes the *directional* alignment between the members based on who is addressing (source) and the addressee (target). The directional alignment is captured over multiple interactions in the meeting.

a) Pairwise Feature Construction.: For any modality vectors $\mathbf{a}$ and $\mathbf{b}$, we define:

$$\phi(\mathbf{a}, \mathbf{b}) = [\mathbf{a} \,\|\, \mathbf{b} \,\|\, |\mathbf{a} - \mathbf{b}| \,\|\, \mathbf{a} \odot \mathbf{b}], \tag{5}$$

where: - $|\mathbf{a} - \mathbf{b}|$ captures absolute differences, - $\mathbf{a} \odot \mathbf{b}$ captures multiplicative similarity, - $\| \cdot \|$ denotes vector concatenation.

Applying this to role, intention, and text embeddings yields:

$$\begin{aligned} \mathbf{z}_{s\to t} = {} & \phi(\mathbf{role}_s, \mathbf{role}_t) \,\| \\ & \phi(\mathbf{intent}_s, \mathbf{intent}_t) \,\| \\ & \phi(\mathbf{utter}_s, \mathbf{utter}_t) \end{aligned} \tag{6}$$

a vector in $\mathbb{R}^{d_z}$, where $d_z = 4(d_r + d_{da} + d_u)$.

b) Interaction Modeling.: A non-linear MLP processes the pairwise features:

$$\mathbf{h}_{s\to t} = \mathrm{MLP}_{\text{int}}(\mathbf{z}_{s\to t}), \tag{7}$$

which learns higher-order relational patterns between the two speakers.

c) Gating Mechanism.: To suppress noisy or uninformative alignment signals, we apply a learned gate:

$$g_{s\to t} = \sigma(\mathrm{MLP}_{\text{gate}}(\mathbf{h}_{s\to t}))\,, \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function.

d) Final Alignment Valence.: The gated representation is projected to the hidden dimension:

$$\mathbf{AV}_{s\to t} = \mathrm{LN}(W_{\text{out}}(g_{s\to t} \cdot \mathbf{h}_{s\to t}))\,, \tag{9}$$

producing a directional alignment vector in $\mathbb{R}^{d_h}$, where d_h is the hidden-state dimension. A symmetric computation yields $\mathbf{AV}_{t\to s}$.

C. Dyadic Message Passing

In this component, the previous hidden states (cognitive states) of the two team members in the conversation are integrated with their directional alignment valences. This step models how each speaker influences the other during the dyadic exchange.

Let $\mathbf{h}_s^{(t-1)}$ and $\mathbf{h}_t^{(t-1)}$ denote the previous hidden states of speakers s and t. We compute directional messages:

$$\mathbf{m}_{s\to t} = \mathrm{MLP}_{\text{msg}}\Big([\mathbf{h}_s^{(t-1)} \,\|\, \mathbf{h}_t^{(t-1)} \,\|\, \mathbf{AV}_{s\to t} \,\|\, \mathbf{AV}_{t\to s}]\Big)\,, \tag{10}$$

$$\mathbf{m}_{t\to s} = \mathrm{MLP}_{\text{msg}}\Big([\mathbf{h}_t^{(t-1)} \,\|\, \mathbf{h}_s^{(t-1)} \,\|\, \mathbf{AV}_{t\to s} \,\|\, \mathbf{AV}_{s\to t}]\Big)\,. \tag{11}$$

Here: - $\mathbf{m}_{s\to t}$ encodes how s influences t, - $\mathbf{m}_{t\to s}$ encodes how t influences s.

D. Temporal Update via GRU

Now that we have the speakers' encoded incoming messages, an update to the speakers' cognitive state is necessary, which is performed by a gated recurrent unit (GRU) cell. The GRU Cell integrates the speaker's previous hidden state with the incoming message to produce an updated mental state for each team member. This mechanism allows the model to accumulate dyadic conversational history over time.

$$\mathbf{h}_s^{(t)} = \mathrm{GRU}(\mathbf{m}_{s\to t}, \mathbf{h}_s^{(t-1)}), \mathbf{h}_t^{(t)} = \mathrm{GRU}(\mathbf{m}_{t\to s}, \mathbf{h}_t^{(t-1)}). \tag{12}$$

Here: - $\mathbf{h}_s^{(t)}$ is the updated hidden state of speaker s, - $\mathbf{h}_t^{(t)}$ is the updated hidden state of speaker t.

E. Alignment Classification

This is the final component of the model, responsible for predicting the alignment category for each dyadic event in the meeting. The cognitive states of the team members in the dyadic event and their respective directional alignment valence are combined to predict the alignment category, as shown in the equation below:

$$\mathbf{z}_{\text{pair}} = [\mathbf{h}_s^{(t)} \,\|\, \mathbf{h}_t^{(t)} \,\|\, \mathbf{AV}_{s\to t} \,\|\, \mathbf{AV}_{t\to s}] \in \mathbb{R}^{4d_h}. \tag{13}$$

A final MLP produces the alignment label:

$$\hat{y} = \mathrm{softmax}(W_c \mathbf{z}_{\text{pair}} + b_c), \tag{14}$$

where $\hat{y}$ is the predicted alignment class.

This formulation enables the model to jointly leverage temporal dynamics, speaker cognitive states, and relational alignment to classify alignment behaviour in conversational interactions.

IV. Results and Discussion

A. Dataset

We evaluate our framework on dyadic interactions extracted from the AMI Meeting Corpus [6], a widely used benchmark dataset for multi-party conversational analysis that contains 100 hours of meeting recordings, including both natural and scenario-driven meetings. The corpus captures naturalistic group interactions, making it particularly suitable for studying conversational dynamics, intention modelling, and multi-speaker behaviour in real-world settings. For our experimental setup, we use only scenario-driven meetings, and since the framework models dyadic exchanges, multi-party meetings are decomposed into ordered adjacency pairs (s, t). For each event, the role, intention, and corresponding utterance text are mapped for both the source and target speakers. All the meetings in the corpus are split into four sessions: a, b, c, and d. We consider each session an independent meeting in our study. Only conversations with complete annotations are retained for training and evaluation.

a) Data Preprocessing: While analysing the dataset, we observed that certain dialogue act categories, such as *Fragment*, *Stall*, and *Be-Negative*, contained very limited information and were therefore removed. Each of the adjacency pairs found in the AMI Corpus is labelled with a conversational intent class as follows: *Support/Positive Assessment*, *Partial agreement/support*, *Objection/Negative Assessment*, and *Uncertain*. While analysing the data, we observed a significant class imbalance. Out of 24,738 adjacent pair samples, *Support/Positive Assessment* accounts for 20,851 instances (84.29%), followed by *Objection/Negative Assessment* with 2,038 (8.24%), *Uncertain* with 1,240 (5.01%), and *Partial agreement/support* with only 609 instances (2.46%). Thus, the dataset is heavily dominated by the support class, with limited representation of the other categories, particularly *Partial agreement/support*.

We ran an overlap analysis between Support/Positive Assessment and Partial agreement/support, which showed substantial feature-space overlap (ROC-AUC = 0.764, silhouette = -0.023, centroid cosine similarity = 0.907), indicating that Partial agreement/support was neither a stable nor a distinct category and was frequently absorbed into the support class. As a result, the Partial agreement/support label was removed before data splitting.

To create non-overlapping *train*, *validation*, and *test* sets at the meeting level, the meetings were split into train/validation/test (80/10/10) using a meeting-level multi-stratified split, which yielded 106 meetings in the train set, containing 19297 adjacent pairs; 17 meetings in the validation set, containing 2407 pairs; and 15 meetings in the test set, containing 2408 pairs.

B. Experimental Setup

All models are trained using the Adam optimiser [36] with weight decay ($1e^{-5}$), and the learning rate ($1e^{-4}$) is selected via validation search. A learning rate scheduler (ReduceLROnPlateau), a Keras callback class with mode="min", factor=0.5, and patience=2, is employed to adaptively reduce the learning rate based on validation performance. The text encoder is a frozen BERT model, and only the projection layer is updated during training. Role and intention embeddings are learned jointly with the rest of the model. Hidden states are reset at the start of each meeting to ensure temporal consistency within the meeting, but not across meetings. In TGN-SMM, all hidden states (d_h, d_r, d_{da}, d_u) are set to 128.

To address class imbalance, training minimises a class-weighted cross-entropy loss. Let n_c denote the number of samples for class c. The class weights are computed as:

$$w'_c = \frac{1}{\sqrt{n_c}}, \quad w_c = \frac{w'_c}{\sum_{k=1}^{C} w'_k} \cdot C, \tag{15}$$

where C is the total number of classes, the final objective is given by:

$$\mathcal{L} = -\sum_{c=1}^{C} w_c \, y_c \log \hat{y}_c, \tag{16}$$

TABLE I
PERFORMANCE COMPARISON ON THE AMI DATASET. CLASS-LEVEL METRICS (PRECISION, RECALL, F1) ARE REPORTED ALONG WITH OVERALL METRICS. BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Model	Class	Prec.	Rec.	F1	Acc.	F1 (macro)	Prec. (macro)	Rec. (macro)	F1 (weighted)	Prec. (weighted)	Rec. (weighted)
TGN-SMM (Ours)	Objection/Negative	0.5167	**0.4408**	**0.4757**	0.8746	**0.6422**	**0.6584**	**0.6284**	**0.8710**	**0.8682**	0.8746
	Support/Positive	**0.9234**	0.9403	0.9317							
	Uncertain	**0.5351**	0.5041	**0.5191**							
Static SMM	Objection/Negative	**0.5789**	0.2085	0.3066	0.8708	0.5843	0.6521	0.5782	0.8556	0.8571	0.8708
	Support/Positive	0.9080	**0.9557**	0.9312							
	Uncertain	0.4694	**0.5702**	0.5149							
Speaker-GRU	Objection/Negative	0.5425	0.3934	0.4560	**0.8767**	0.6226	0.6516	0.6020	0.8702	0.8663	**0.8767**
	Support/Positive	0.9211	0.9499	**0.9353**							
	Uncertain	0.4912	0.4628	0.4766							

where y_c is the ground-truth label and $\hat{y}_c$ is the predicted probability for class c.

a) Baselines: We evaluate the proposed TGN-SMM against variants that progressively incorporate its key components. *Static SMM* removes temporal dynamics, computing speaker representations solely from multimodal embeddings without message passing or GRU updates. *Speaker-GRU* introduces temporal modelling via a GRU-based update of speaker states, but excludes alignment valence and dyadic message passing. *TGN-SMM (ours)* represents the full model, combining multimodal speaker embeddings, directional alignment valence, dyadic message passing, and temporal updates.

C. Results

Table I shows the performance of the proposed TGN-SMM model compared to both the static and temporal SMM baselines. The results demonstrate the effectiveness of TGN-SMM, which achieves the highest macro-averaged F1 score and outperforms the other baselines. This indicates that our proposed model better captures the internal team dynamics used to infer the alignment states against which performance is compared. TGN-SMM also achieves the highest weighted F1 score, reflecting strong performance on the dataset distribution. The results also demonstrate that integrating temporal dynamics through message passing and GRU-based updates significantly enhances the model's ability to capture conversational dynamics over time. Compared with the Speaker-GRU, the results indicate that alignment-aware valence learning yields a better class balance rather than solely improving majority-class predictions.

The results suggest that the primary challenge for all SMMs in this task is effectively modelling minority and semantically subtle classes. While all models perform quite well on the majority class *Support/Positive*, TGN-SMM shows improved performance on the more challenging *Objection/Negative* and *Uncertain* classes. This would suggest that the alignment valence and dyadic message passing enable the model to discriminate better subtle interaction patterns that are often overlooked by purely temporal or static representations. These classes might be more dependent on conversational dynamics than on semantic cues in the utterances, which may also explain why the static SMM struggles, particularly in recall.

Overall, the results imply that the combination of temporal dynamics and structured interpersonal interaction modelling is particularly beneficial for handling under-represented and context-dependent dialogue acts, which are the core difficulty in this dataset. The results indicate that the combination of modalities used in TGN-SMM yields more robust and balanced performance for multi-party dialogue understanding.

V. CONCLUSION

In this work, we proposed TGN-SMM, a temporal graph-based framework for modelling interpersonal team dynamics and shared cognitive states in multi-party conversations. The experimental results on the AMI Meeting Corpus demonstrated that TGN-SMM outperformed strong baselines on both macro and weighted metrics, indicating improved performance across both minority and majority classes, with Speaker-GRU having the best overall accuracy. The results highlight the importance of explicitly modelling interaction alignment and temporal dependencies in capturing conversational dynamics. Overall, the proposed approach provides a principled way to model evolving shared mental representations in group interactions, offering a promising direction for future work in multi-party dialogue understanding, collaborative behaviour analysis, and socially aware AI systems.

In Future Work, we would like to extend our work by developing a concrete human-agent deployment scenario that leverages representations from the TGN-SMM to support the agent's decision-making and proactive assistance in achieving shared team goals.

ACKNOWLEDGMENT

The authors acknowledge the use of AI-based tools for language editing and improving the clarity of the manuscript. All technical content, analysis, and conclusions remain the sole responsibility of the authors.

This work was carried out within the projects "AI CoWorking Lab Network (AICOLABS)" and "Resilient human-robot collaboration in a mixed-skill environment (ENABLING)", financed by the European Regional Development Fund (ERDF), Saxony-Anhalt. AICOLABS is co-funded by the Federal state of Saxony-Anhalt and the EU (ESF+), Funding Nr. ZS/2024/01/183362.

References

[1] A. Nowak, P. Lukowicz, and P. Horodecki, "Assessing artificial intelligence for humanity: Will ai be the our biggest ever advance? or the biggest threat [opinion]," *IEEE Technology and Society Magazine*, vol. 37, no. 4, pp. 26–34, 2018.

[2] R. Iftikhar, Y.-T. Chiu, M. S. Khan, and C. Caudwell, "Human–agent team dynamics: A review and future research opportunities," *IEEE Transactions on Engineering Management*, vol. 71, pp. 10 139–10 154, 2023.

[3] E. Mathieu John, F. Goodwin Gerald, S. Heffner Tonia, and C.-B. J. A. Salas Eduardo, "The influence of shared mental models on team process and performance," *Journal of Applied psychology*, vol. 85, no. 2, pp. 273–283, 2000.

[4] R. W. Andrews, J. M. Lilly, D. Srivastava, and K. M. Feigh, "The role of shared mental models in human-ai teams: a theoretical review," *Theoretical Issues in Ergonomics Science*, vol. 24, no. 2, pp. 129–175, 2023.

[5] C. M. Jonker, M. B. Van Riemsdijk, and B. Vermeulen, "Shared mental models: A conceptual analysis," in *International workshop on coordination, organizations, institutions, and norms in agent systems*. Springer, 2010, pp. 132–151.

[6] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International workshop on machine learning for multimodal interaction*. Springer, 2005, pp. 28–39.

[7] G. Hoffman and C. Breazeal, "Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team," in *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, 2007, pp. 1–8.

[8] S. Mohammed, L. Ferzandi, and K. Hamilton, "Metaphor no more: A 15-year review of the team mental model construct," *Journal of management*, vol. 36, no. 4, pp. 876–910, 2010.

[9] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran, "Interactive team cognition," *Cognitive science*, vol. 37, no. 2, pp. 255–285, 2013.

[10] M. Chatterjee, B. Ensan, A. Wendemuth, and A. Al-Hamadi, "Think like a team: Graph-based representation of shared mental models in human-agent collaboration," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2026, Tagungsband der 37. Konferenz*, G. Wirsching, Ed. TUDpress, Dresden, Mar. 2026, pp. 151–158. [Online]. Available: https://www.essv.de/pdf/2026_151_158.pdf

[11] J. A. Cannon-Bowers, E. Salas, and S. Converse, "Shared mental models in expert team decision making," *Individual and group decision making: Current issues*, vol. 221, pp. 221–46, 1993.

[12] B.-C. Lim and K. J. Klein, "Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy," *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, vol. 27, no. 4, pp. 403–418, 2006.

[13] S. Mohammed, K. Hamilton, R. Tesler, V. Mancuso, and M. McNeese, "Time for temporal team mental models: Expanding beyond "what" and "how" to incorporate "when"," *European Journal of Work and Organizational Psychology*, vol. 24, no. 5, pp. 693–709, 2015.

[14] C. M. Santos, A. M. Passos, and S. Uitdewilligen, "When shared cognition leads to closed minds: Temporal mental models, team learning, adaptation and performance," *European Management Journal*, vol. 34, no. 3, pp. 258–268, 2016.

[15] C. M. Santos, S. Uitdewilligen, and A. M. Passos, "A temporal common ground for learning: The moderating effect of shared mental models on the relation between team learning behaviours and performance improvement," *European Journal of Work and Organizational Psychology*, vol. 24, no. 5, pp. 710–725, 2015.

[16] P. Van den Bossche, W. Gijselaers, M. Segers, G. Woltjer, and P. Kirschner, "Team learning: building shared mental models," *Instructional science*, vol. 39, no. 3, pp. 283–301, 2011.

[17] L. A. DeChurch and J. R. Mesmer-Magnus, "Measuring shared team mental models: A meta-analysis." *Group dynamics: Theory, research, and practice*, vol. 14, no. 1, p. 1, 2010.

[18] J. C. Gorman, N. J. Cooke, P. G. Amazeen, and S. Fouse, "Measuring patterns in team interaction sequences using a discrete recurrence approach," *Human Factors*, vol. 54, no. 4, pp. 503–517, 2012.

[19] R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters, "Modelling and detecting decisions in multi-party dialogue," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008, pp. 156–163.

[20] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching dbn," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 7, pp. 1303–1314, 2008.

[21] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[22] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation," in *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009, pp. 7–14.

[23] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings the ami and amida projects," in *2008 Hands-Free Speech Communication and Microphone Arrays*. IEEE, 2008, pp. 115–118.

[24] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and W. B. Dolan, "A persona-based neural conversation model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 994–1003.

[27] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 154–164.

[28] V. W. Feng and G. Hirst, "A linear-time bottom-up discourse parser with constraints and post-editing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 511–521.

[29] W. Hu, Z. Chan, B. Liu, D. Zhao, J. Ma, and R. Yan, "Gsn: A graph-structured network for multi-party dialogues," *arXiv preprint arXiv:1905.13637*, 2019.

[30] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," *arXiv preprint arXiv:2006.10637*, 2020.

[31] S. Kumar, X. Zhang, and J. Leskovec, "Predicting dynamic embedding trajectory in temporal interaction networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 1269–1278.

[32] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: a survey," *ACM computing surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[33] R. Trivedi, H. Dai, Y. Wang, and L. Song, "Know-evolve: Deep temporal reasoning for dynamic knowledge graphs," in *international conference on machine learning*. PMLR, 2017, pp. 3462–3471.

[34] J. Eisenstein and R. Barzilay, "Bayesian unsupervised topic segmentation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 334–343.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.