Enhancing Gaze Prediction in Multi-Party Conversations via Speaker-Aware Multimodal Adaptation

Meng-Chen Lee
Department of Computer Science
University of Houston
Houston, Texas, USA
mlee45@uh.edu

Zhigang Deng
Department of Computer Science
University of Houston
Houston, Texas, USA
zdeng4@central.uh.edu

ABSTRACT

Modeling gaze patterns in multiparty conversations is crucial to build socially-aware dialogue agents and humanoid robots. However, existing approaches typically rely on visual data or focus on dyadic settings. We propose a novel framework for social attention modeling — predicting gaze directions from linguistic and speaker cues alone, without direct visual input. We introduce SAT5, a speaker-aware adaptation of the T5 language model, pre-trained using multi-task objectives that capture both span corruption and speaker state modeling. Using a new dataset of three-party face-to-face conversations with synchronized speech, gaze, and motion capture data, we demonstrate that SAT5 significantly outperforms both pretrained and RNN-based baselines in predicting gaze targets. Our findings highlight the importance of conversational structure and speaker dynamics in modeling social attention, and offer a strong foundation for gaze-aware multimodal systems.

CCS CONCEPTS

• **Human-centered computing → Collaborative interaction**; **Empirical studies in HCI**.

KEYWORDS

multi-party conversations; human communication dynamics; multimodal interaction; non-verbal gesture; human-human interaction

ACM Reference Format:
Meng-Chen Lee and Zhigang Deng. 2025. Enhancing Gaze Prediction in Multi-Party Conversations via Speaker-Aware Multimodal Adaptation. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3716553.3750755

1 INTRODUCTION

In recent years, dialogue agents, including interactive robots and virtual agents, have become increasingly prevalent across various domains, from customer service and healthcare to education and entertainment [17]. As these systems continue to evolve, their ability to facilitate natural and engaging communication with humans has become a key area of research. While early dialogue agents relied primarily on rule-based or scripted interactions, advances in artificial intelligence, particularly in natural language processing (NLP) and multimodal learning, have enabled more sophisticated and context-aware conversational agents. However, achieving seamless and intuitive human-agent interactions remains an ongoing challenge, particularly in multi-party conversation settings where interaction dynamics are more complex.

Most of existing research on dialogue systems has focused on dyadic (two-party) interactions, where conversational structure is relatively straightforward. However, real-world conversations often involve multiple interlocutors, such as in meetings, group discussion, and social gatherings. Multi-party dialogues introduce additional complexities, including managing conversational flow, handling speaker transitions, coordinating turn-taking, and interpreting overlapping speech. These challenges make multi-party interactions a more realistic yet demanding research problem, requiring models that can process dynamic conversational cues and predict interaction patterns in real time. Addressing these challenges is crucial for advancing the capabilities of conversational AI, enabling agents to participate more effectively in group discussion and collaborative environments.

Beyond linguistic content, nonverbal cues, such as gestures, prosody, facial expressions, and gaze, play an essential role in structuring and regulating interactions [20, 21]. These cues provide additional layers of meaning beyond spoken words, helping participants establish engagement, emphasis, and interpersonal alignment. Various speech-driven motion generation approaches have been explored, including models for full-body gestures [35] and hand gestures [12]. While these modalities contribute to expressive communication, gaze behavior holds particular importance in multi-party dialogue settings. Unlike other nonverbal signals, gaze explicitly guides attention, signals turn-taking intentions, and helps synchronize interactions. As such, understanding and modeling gaze behavior is essential for creating more responsive and context-aware dialogue agents.

Gaze serves multiple communicative functions, particularly in regulating turn-taking [7, 19], and conveying emotions, intentions, and cognitive states [1, 8, 30]. For example, speakers may avert their gaze to indicate cognitive processing or signal an intention to retain their turn [7]. Conversely, listeners often use gaze to signal engagement, monitor speaker cues, and anticipate conversational shifts. These subtle but highly informative gaze patterns are fundamental to maintaining coherent and fluid conversations, particularly in multi-party settings where participants must continually navigate shifting attention, speaker roles, and social cues.

ICMI '25, October 13–17, 2025, Canberra, ACT, Australia

ACM ISBN 979-8-4007-1499-3/2025/10
https://doi.org/10.1145/3716553.3750755

Moreover, gaze behavior is not only reactive but also modulated by speech content. Patterns, such as gaze convergence during emphasis or divergence during hedging, can reflect underlying communicative intent and facilitate mutual understanding [19, 26]. This tight coupling between speech and gaze highlights the importance of modeling their interplay to support naturalistic interaction. Given its significance, accurately predicting and generating gaze behavior in human-agent dialogue can greatly enhance an agent's naturalness, responsiveness, and social intelligence. By incorporating gaze-aware models, virtual agents and robots can more effectively manage turn-taking, sustain engagement, and adapt to conversational context, resulting in more intuitive and effective interactions in multi-party scenarios.

In this work, building upon the T5 architecture [29], we introduce *Speaker-Aware* T5 (called SAT5), a novel adaptation of pretrained language models designed for speaker-aware representation learning. The main contributions of this work can be summarized as follows: (1) We design a circular coordinate system to embed speaker identity, enabling generalization across varying numbers of speakers in multi-party settings. (2) We extend the T5 architecture with speaker-aware embeddings and a pretraining strategy that jointly models speaker identities and span corruption, allowing the model to capture dynamic speaker transitions and conversational structures beyond linguistic features. (3) We apply our adapted model to the downstream task of gaze prediction, leveraging multimodal conversational data to accurately forecast eye gaze targets in group interactions.

2 RELATED WORK

Speaker-Aware Approaches. Many research efforts have focused on developing neural networks and pre-training speaker-aware language models to enhance multi-party conversation understanding. From the perspective of recurrent neural networks (RNNs), researchers introduced the Multi-Party Conversation (MPC) corpus, enabling tasks such as addressee identification and response selection using different RNN architectures [24, 27]. Similarly, Le et al. [18] proposed the who-to-whom (W2W) architecture, which integrates speaker and listener-specific gated recurrent units (GRUs), namely the Speaker GRU (SGRU) and Listener GRU (LGRU), to model conversational dynamics more effectively.

Recently, transformer-based architectures have demonstrated significant advancements in speaker-aware modeling for multiparty dialogue. Gu et al. [9] introduced Speaker-Aware BERT (SA-BERT), an extension of the BERT model that incorporates speaker embeddings to enhance contextual representations in dialogue systems. Later, Gu et al. [10] further extended this work by employing a multitask learning approach to improve the overall comprehension of multi-party conversations across various downstream tasks.

These models have proven effective in capturing discourse coherence within textual dialogues by leveraging self-supervised learning objectives. Additionally, they have demonstrated the robustness of RNN-based architectures and the generalization capabilities of pre-trained language models. However, a critical limitation of these approaches is that they primarily operate on textual data alone, overlooking the explicit temporal structures present in spoken dialogues, such as pauses, silences, and interruptions. In real-world multi-party conversations, these temporal features play a crucial role in structuring interactions, regulating turn-taking, and managing speaker transitions. The absence of such cues limits the effectiveness of text-only models in capturing the real-time dynamics of multi-party conversations.

Gaze Prediction. Although no prior work has explicitly addressed gaze prediction in multi-party conversations, several related studies have explored gaze prediction and analysis in different contexts, offering valuable insights that inform our approach to modeling gaze behavior in dynamic, multi-party interactions.

A significant body of research has focused on gaze prediction in reading tasks. Agarwal and Chatterjee [2] demonstrated that linguistic features, such as shallow lexical characteristics, token rarity, and token interactions, contribute to improved gaze predictions during reading, emphasizing the role of linguistic cues in gaze modeling. Additionally, studies such as [22, 25, 33] have showcased the effectiveness of state-of-the-art transformer models in gaze prediction, further highlighting the benefits of deep contextualized representations. Despite these advancements, existing gaze prediction models primarily focus on reading-based tasks.

Human–robot interaction (HRI) research has tackled gaze prediction from visual inputs. For example, Saran et al. [31] developed a deep-learning system that predicts referential and mutual gaze using robot-mounted cameras, combining head-direction estimation with object saliency cues to enable real-time gaze following [4]. In computer vision, Tu et al. [32] introduced HGTTR, an end-to-end transformer-based model that simultaneously detects human head locations and corresponding gaze targets in complex scenes, achieving strong results on benchmark datasets. Subsequent work such as TransGOP [34] further applied transformer-based object detectors to model long-range head–gaze target relationships, while MGTR [11] focused on mutual gaze detection in social imaging using transformer architectures. Although these visual-scene and HRI approaches demonstrate robust gaze-target prediction, they concentrate on dyadic or static image settings and do not incorporate conversational or linguistic context.

Our work bridges this gap by targeting gaze target prediction in multi-party conversation, a dynamic, interactive domain that merges insights from reading-based gaze prediction and visual-scene gaze modeling. To our knowledge, this is the first study to combine transformer-based gaze-target prediction with linguistic and speaker-role context in a realistic, multi-person interaction scenario.

3 DATASETS

We utilized two multi-party conversation datasets to train and evaluate our model: (1) the AMI corpus, and (2) an InHouseDataset, described below.

3.1 AMI Corpus

The AMI Meeting Corpus [16] is a richly annotated multi-modal dataset of around 100 hours of real and scenario-based business meetings recorded at multiple sites, capturing participants through tables and personal microphones, as well as panoramic and individual cameras. The collection includes both spontaneous discussion and structured "design team" meetings where participants assume

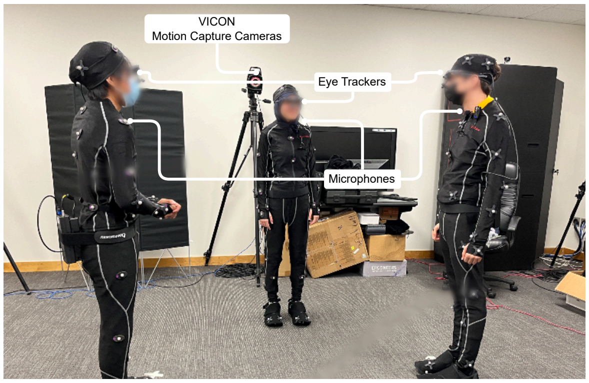

Figure 1: A snapshot of our three-party conversational motion data acquisition experiment.

specific roles in a product-design task. In our study, we used this dataset exclusively for the initial training stage, as detailed in Section 4.

3.2 InHouseDataset

To construct an InHouseDataset, we recruited 21 volunteers from a university campus and randomly assigned them into seven groups of three. None of the participants was acquainted before the experiment. Among them, 12 were men and 9 were women, aged between 20 and 30. The majority (70%) had a background in computer science, and all were native English speakers. The data capture was approved by the Institutional Review Boards.

Each group participated in three to five recorded sessions of three-party conversations. The session durations for the seven groups were 46 minutes 18 seconds (46'18"), 44'47", 46'17", 42'49", 46'31", 44'27", and 48'19", totaling 319 minutes and 28 seconds of recorded data. For each participant, we employed Ergoneers Dikablis Glass 3 for eye tracking, a wireless microphone for speech recording, and an optical motion capture suit to accurately capture movements of the head, hands, torso, and lower body.

During data collection, participants were instructed to maintain fixed positions, forming an equilateral triangle with approximately 1 meter of separation between them. This spatial configuration aligns with setups used in prior studies on three-party conversations [3, 6, 13–15] and is similar to the pentagonal arrangement for five-party interactions [26]. Figure 1 offers a snapshot of the actual data acquisition process in action.

Participants engaged in free discussion on topics of their choice throughout the recording sessions. To ensure precise synchronization across multiple data sources, we used a clapperboard at the beginning of each session. As a result, all collected data —including 3D motion capture, eye tracking, and speech — were temporally aligned and subsequently downsampled to 30 frames per second for consistency.

Inspired by previous research [14], we estimated an interlocutor's Direction-of-Focus (DFoc) by integrating both torso-head orientation and the eye gaze direction vector (V_E). Specifically, given the 3D rotational representations of the hips (R_{hips}), spine (R_S), and head (R_H), along with the global position of the hips (G_h) and the offsets of the spine (O_S), head (O_H), and eyes (O_E), we defined DFoc as follows:

$$DFoc = G_h R_{hips} O_S R_S O_H R_H O_E V_E. \tag{1}$$

To ensure the robustness of this computation, we addressed missing or anomalous values in V_E by employing piecewise shape-preserving cubic interpolation. Building upon the DFoc estimation, we further introduced a higher-order feature, Focus of Attention (FoA), which identifies the interlocutor an individual is attending to at a given frame.

To formally represent FoA, we defined a Gaze Target Vector (GTV) as a 1×3 ternary vector for each frame. This vector effectively encodes the attention focus for all three interlocutors. Specifically, the i-th component of GTV corresponds to the index of the interlocutor being attended to by the i-th individual at that frame.

3.3 Segmentation

In both datasets, we sampled data at 30 frames per second (fps), resulting in a frame duration of approximate 33 milliseconds, ensuring alignment across all modalities. To determine the optimal window size for training, step size for augmentation, and discrete chunk size for maintaining a uniform representation, we analyze the average durations of speech and silence.

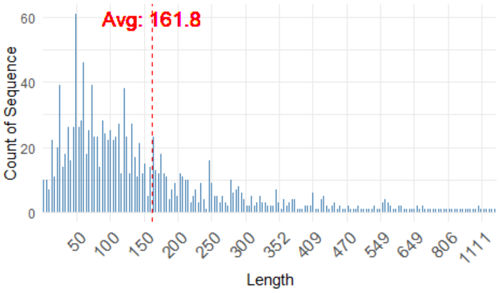

Figure 2: Count and average length of speaking sequences.

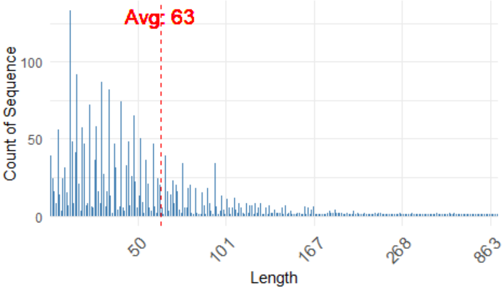

Figure 3: Count and average length of silence sequences.

Instead of a token-based approach, we adopt a time-based prediction method to effectively capture gaps and silences between words

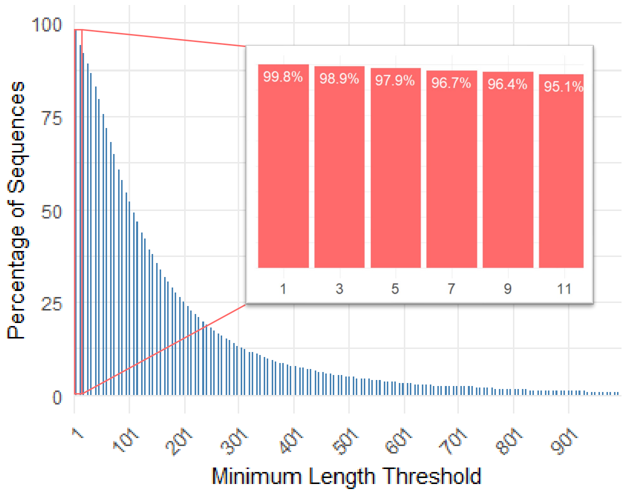

Figure 4: Percentages of sequences that are longer than the thresholds.

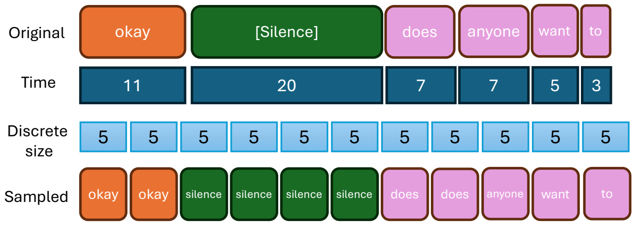

Figure 5: Data sampled to a discrete chunk size 5 by assigning the majority behavior to ensure the uniform representation across all modalities.

and sentences. Our analysis in Figure 2 and Figure 3 shows that the average speaking duration before a speaker switch is 162 frames (5.4 seconds), while the average silence between speech segments is 63 frames (2.1 seconds). Based on this, we set the window size to 450 frames (15 seconds) to capture sufficient conversational context, including speech, silence, and speaker transitions. The step size is set to 150 frames (5 seconds) to ensure an adequate number of training instances while minimizing excessive overlap and redundancy.

Additionally, as shown in Figure 4, a discrete chunk size of 5 frames covers 97.9% of utterances, making it a suitable choice for determining prediction frequency while mitigating erratic gaze shifts. This approach reduces training redundancy while maintaining prediction accuracy.

Furthermore, Figure 5 illustrates the final downsampled representation. To ensure consistency across modalities, we aggregate words and speaker identities within each discrete chunk by assigning the most frequently occurring feature in that interval. This strategy is also applied to gaze signals, which helps reduce abrupt shifts and enhances the robustness of the representation. The colors shown in Figure 5 denote different speaker states—either active speakers or silence. The implementation details of speaker embedding are further elaborated in Section 4.2.1.

4 SPEAKER STATE ADAPTATION

4.1 Input Representation

We define each data instance as a sequence of pairs $\{(T_c, S_c)\}_{c=1}^{C}$, where T_c denotes the text embedding of the c-th discrete chunk and S_c is the corresponding speaker embedding. Specifically, T_c represents the embedded representation of the textual content in the c-th chunk, obtained by passing the corresponding token sequence through the embedding layer of T5. This transforms discrete token indices into continuous vectors within the T5 model's representational space, capturing the chunk's semantic content. The embedding S_c encodes the speaker identity using a circular coordinate system, where each speaker is assigned an angular position on a unit circle and projected into the same embedding space as T_c through a learnable transformation function. The combined chunk representation is formed by chunk-wise addition of T_c and S_c, and is subsequently fed into the T5 encoder.

In our setup, we define an utterance window of 450 frames, segmented into $C = 90$ discrete chunks, each corresponding to 5 frames of audio and aligned text. This chunk-level formulation enables us to inject speaker context at a fine-grained temporal resolution while maintaining alignment with downstream gaze and dialogue modeling objectives.

Our primary goal is to develop a speaker-aware pre-trained language model that captures both semantic and speaker-specific information. Given an instance, the model is expected to output embedding vectors for all 90 discrete chunks, enriching them with the structure of the speaker identity. These embeddings can then be fine-tuned on various downstream tasks, providing a flexible framework that leverages both linguistic content and speaker context.

4.2 Speaker-Aware T5 (SAT5)

In this work, we build upon the T5 architecture [29] to incorporate speaker-awareness. We refer to our extended model as Speaker-Aware T5 (SAT5). The original T5 model is pre-trained on a large-scale text corpus, capturing general linguistic representations. However, in order to adapt T5 to our speaker-focused task, we perform domain adaptation on in-domain corpora.

4.2.1 *Incorporating Speaker Embeddings.*

To account for speaker information in each utterance, we introduce a *circular speaker embedding system* into the standard T5 input representation (see Figure 7). Specifically, each discrete chunk of text T_c is paired with a corresponding speaker embedding S_c, which encodes the speaker identity as an angle $\theta \in [0, 2\pi)$ on a unit circle, as illustrated in Figure 6. Given N total active speakers in the conversation and an additional label for silent segments (i.e., no speaker), we allocate $N + 1$ angular coordinates on the circle. Each label n is assigned an angular coordinate as follows:

$$\theta^n = \frac{2\pi n}{N+1}, \qquad n \in 0, 1, \ldots, N \tag{2}$$

Here, $n = 1, \ldots, N$ correspond to actual speakers, while $n = 0$ denotes silent or non-speech segments. This results in a spatially uniform distribution of all speaker and non-speaker states along the circle.

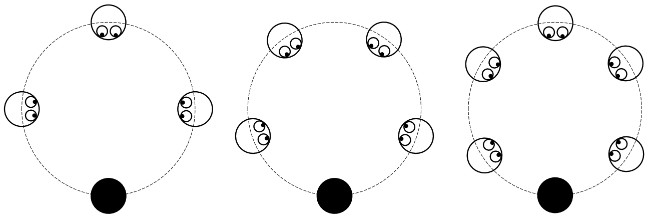

Figure 6: Illustration of the polarized speaker distribution assumption in settings with three, four, and five participants. In speaker state modeling, each speaker is assigned a fixed position on a unit circle, and an additional position—marked by a black circle—is reserved for silent or non-speech segments. The model predicts an angle on this circle, and the speaker state is determined by selecting the label (either a speaker or silence) whose assigned position is the closest to the predicted angle.

The angle θ^n is passed through a learnable transformation function E_s that maps it to the same dimension as T5 token embeddings:

$$S_c = E_s(\theta^n), \quad S_c \in \mathbb{R}^d, \tag{3}$$

where d is the T5 embedding dimension. The resulting speaker embedding S_c is then added to the token embeddings of chunk T_c, forming the combined input to the T5 encoder:

$$\tilde{T}_c = T_c + S_c \tag{4}$$

This polar encoding offers two primary advantages. First, it establishes a consistent reference frame across sessions by uniformly distributing roles (including silence) in a structured geometric space, which helps the model learn speaker transition patterns and turn-taking dynamics. Second, although the layout assumes equal angular spacing for simplicity, the framework is flexible and can accommodate non-uniform or dynamic speaker arrangements. In practice, the angular positions can be re-parameterized post hoc to reflect real-world speaker locations or spatial asymmetries without changing the model architecture.

The enriched input embeddings $\tilde{T}_c$ are processed by the T5 encoder to generate contextual representations that integrate both semantic and speaker-aware cues, which are then used by the decoder for downstream tasks such as gaze prediction or turn transition modeling.

4.2.2 Pre-training on the AMI Corpus. We conduct pre-training using the AMI corpus, which contains conversations among varying numbers of participants. To handle different numbers of speakers, we design dynamic speaker prediction projection layers that automatically adapt to the number of speakers present in a given conversation. This flexibility allows SAT5 to generalize across multiparty interactions with varying participant counts.

4.3 Multi-Task Learning Objectives

4.3.1 Speaker State Modeling. Identifying the speaker of each utterance is crucial in multi-party conversational scenarios. To enable the model to learn this, we propose a speaker state modeling task. During pre-training, 15% of the speaker embeddings S_c are randomly selected and replaced with a special `[Mask]` embedding to prevent information leakage. The model must then predict the original speaker identity based on the surrounding context.

The output of the model in this task is a scalar angle $\hat{\theta}_c \in [0, 2\pi)$ that corresponds to the predicted speaker for the c-th masked chunk. Since speakers are encoded as angles on the unit circle, it is essential to compare angular predictions in a way that respects circular geometry. Direct subtraction of angles may lead to large errors near the boundary (e.g., comparing 0 and $2\pi - \epsilon$), so we compute a circular distance between the predicted angle $\hat{\theta}_c$ and the ground truth angle θ_c using the following formula,

$$\Delta\theta = \left(\left(\hat{\theta}_c - \theta_c + \pi\right) - 2\pi \cdot \left\lfloor \frac{\hat{\theta}_c - \theta_c + \pi}{2\pi} \right\rfloor \right) - \pi, \tag{5}$$

which wraps angular differences into the interval $[-\pi, \pi)$. It is equivalent to computing the remainder when $(\hat{\theta}_c - \theta_c + \pi)$ is divided by 2π, and then shifting it back by π. This ensures that angular differences are computed along the shortest path on the circle and that the error is invariant to how angles wrap around at 0 and 2π.

The speaker state modeling loss is then defined as the mean squared error over these circular distances:

$$\mathcal{L}_{\text{ssm}} = \frac{1}{C} \sum_{c=1}^{C} (\Delta\theta)^2 . \tag{6}$$

This formulation provides a smooth and rotation-invariant training signal, allowing the model to learn fine-grained speaker representations on the unit circle, without being affected by discontinuities at angular boundaries.

4.3.2 Span Corruption. To further enhance the model's ability to capture linguistic dependencies and contextual cues, we adopt the span corruption objective from the original T5 pre-training scheme. This objective involves corrupting certain spans of text (including potential silences or repeated tokens) and then training the model to reconstruct the original text. The input window w is corrupted by masking multiple spans of tokens, replacing them with sentinel tokens (special placeholder tokens like <extra_id_1>, <extra_id_2>, etc.). The model is trained to predict the missing spans in the decoder. The loss function follows a standard autoregressive cross-entropy loss over the sequence of missing tokens in the output:

$$\mathcal{L}_{sc} = -\sum_{i=1}^{|t|} \log P_\theta(t_i | w', t_{<i}). \tag{7}$$

By jointly optimizing $\mathcal{L}_{cp}$ and $\mathcal{L}_{ssm}$ in a multi-task fashion, our SAT5 model internalizes both speaker-centric and linguistic knowledge. This provides a robust foundation for subsequent fine-tuning on various downstream applications. In summary, our proposed SAT5 model integrates speaker embeddings into the T5 architecture and is pre-trained on the AMI corpus to handle conversations with varying speaker configurations. We employ a multi-task learning framework that addresses both speaker state modeling and span corruption. These objectives are jointly optimized by minimizing the sum of their respective losses as:

$$\mathcal{L} = \alpha\mathcal{L}_{ssm} + \mathcal{L}_{sc}. \tag{8}$$

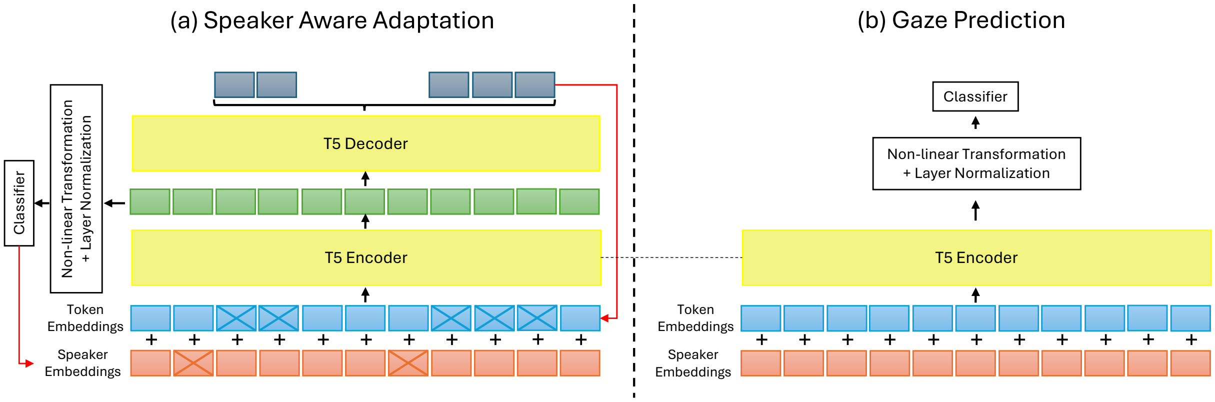

Figure 7: Illustration of the input representations and model architectures: (a) Speaker-Aware Adaptation, where both token and speaker embeddings are masked during pre-training (red arrows indicate cross-entropy computation against ground truth). (b) Gaze Prediction, which uses the pre-trained encoder from (a).

5 GAZE PREDICTION

In this work, without loss of generality, we focus on the task of predicting eye gaze targets in specific *three-party* conversational settings, specifically to determine whether an interlocutor is directing the gaze toward any of other interlocutors or looking elsewhere at any particular moment. That is, the instance for the task is $\{(T_n, S_n, GTV_n)\}_{n=1}^{N}$, with an additional GTV of the n-th discrete chunk of a sequence.

Our proposed gaze prediction framework is built upon the SAT5 encoder model, which has been pre-trained for adaptation as described in Sec. 4. The encoder processes input features extracted from multimodal conversational data, capturing relevant contextual and speaker-specific information necessary for gaze prediction.

To obtain the final gaze predictions, the encoder outputs are passed through a fully connected (FC) layer, which transforms the learned hidden representations into a structured 9-dimensional output space. This output representation corresponds to the predicted gaze directions of three interlocutors, where each interlocutor's gaze is categorized into one of three possible classes G: left interlocutor, right interlocutor, or none. Under the supervised training approach on multi-class classification, we perform cross-entropy between the original GTV and the predicted gaze vector GTV' as:

$$\mathcal{L}_{gtv} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{g=1}^{G} GTV_{n,g}\log(GTV'_{n,g}). \tag{9}$$

6 EXPERIMENTAL SETUP

We adopted the `T5-large` model for all our experiments. During the pre-training of SAT5, the coefficient α for the loss function $\mathcal{L}_{ssm}$ was set to 0.5. For both SAT5 and the gaze prediction model, ReLU was used as the activation function for all non-linear transformations, while softmax was applied to all classification tasks during cross-entropy computation. Optimization was performed using the AdamW optimizer [23]. The training process began with an initial $1,000$ iterations at a learning rate of 1×10^{-4}, followed by $45,000$ iterations with weight decay, starting at 5×10^{-5}. The batch size and dropout rate were set to 32 and 0.15, respectively. All models were trained using a single GeForce RTX 4090 GPU.

6.1 Comparison Methods

To evaluate our approach, we compared it against several baseline models, including both non-pre-training-based and pre-training-based methods. Since no existing models are designed specifically for our task, we adapt the baselines to align with our experimental setting.

Non-pre-training-based models. Previous works, including [18, 24, 27], have explored variations of recurrent neural networks (RNNs), including DRNN, SI-RNN, LSTM, and GRU, which integrate speaker embeddings into conversation modeling. These models have demonstrated effectiveness in capturing speaker awareness within dialogues. However, they rely solely on textual and speaker information, without accounting for temporal gaps or silence.

To establish a baseline for non-pretraining-based approaches, we implemented standard RNN, GRU, and LSTM models. Following previous studies, we utilized 300-dimensional word embeddings pre-trained with *GloVe* [28]. For consistency across models, we set the hidden dimension to 300 and used a two-layer architecture.

Pre-training-based models. BERT [5] is a transformer-based model pre-trained on large-scale text corpora using masked language modeling (MLM) and next sentence prediction (NSP) to learn general language representations. T5 [29] adopts a text-to-text framework, formulating all NLP tasks as sequence-to-sequence problems. For our comparison, we employ these two pre-trained language models without any speaker-aware adaptation.

6.2 Metrics

To assess the effectiveness of our model, we implemented a rigorous evaluation framework designed to ensure both reliability and comprehensiveness. This methodology enables a thorough assessment of the model's performance across diverse scenarios and conversational groups.

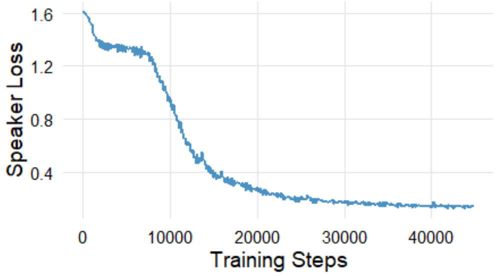

Figure 8: The speaker state modeling loss ($\mathcal{L}_{ssm}$) over training steps, averaged per 100 steps.

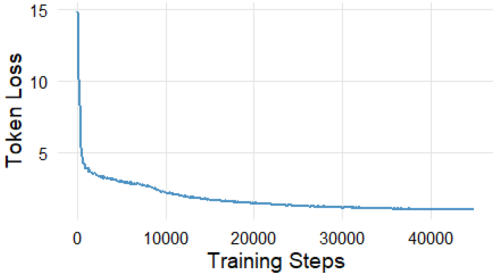

Figure 9: The span corruption loss $\mathcal{L}_{sc}$ over training steps, averaged per 100 steps.

We employed 10-fold cross-validation on the training data, a widely used technique that enhances the robustness of model evaluation. This procedure systematically partitions the dataset into ten equally sized subsets. In each iteration, nine subsets are used for training, while the remaining one serves as the test set. To further prevent overfitting, the training set is internally split into 80% for training and 20% for validation. This ensures that the model is optimized effectively before final evaluation. The cross-validation process is repeated ten times, with each subset serving as the test set once, ensuring a comprehensive and unbiased evaluation.

To quantify performance, we aggregated results across all iterations and computed key evaluation metrics, including precision, recall, and F-score, using a macro-averaging approach. This method assigns equal weights to all classes or conversational behaviors, providing a balanced and holistic assessment of the model's overall effectiveness.

7 RESULTS

Figure 8 and Figure 9 illustrate the effectiveness of our SAT5 model in capturing both speaker representations and text structures, including silences and interruptions. The steady decline in the speaker state modeling loss over training steps indicates that the model effectively learns to differentiate between speakers, adapting to dynamic conversational contexts. Simultaneously, the token loss curve reflects the model's ability to encode textual content while accounting for pauses and disruptions in speech, which are crucial for modeling multiparty interactions in the real world. These results highlight the capability of the SAT5 to integrate both speaker-aware and linguistic features, making it well suited for gaze prediction and other downstream tasks that rely on nuanced conversational dynamics.

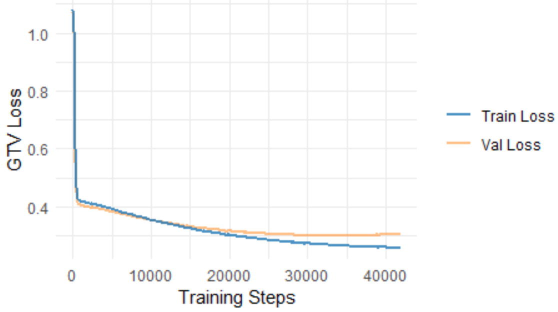

Figure 10: The training and validation losses of GTV Loss ($\mathcal{L}_{GTV}$) over training steps, with average loss computed per 100 steps.

Figure 10 presents the training and validation loss curves for GTV Loss ($\mathcal{L}_{GTV}$) over the course of training, with the average loss computed per 100 steps. The consistent downward trend in both training and validation losses indicates the model's ability to learn gaze target predictions effectively. The minimal gap between training and validation losses suggests that the model generalizes well to unseen data, with no significant overfitting. These results highlight the robustness of our SAT5-based gaze prediction model in capturing multimodal conversational dynamics.

7.1 Quantitative Performance

7.1.1 Speaker State Modeling. Table 1 presents the quantitative evaluation of speaker state modeling during the pretraining stage. The results demonstrate that our SAT5 model effectively captures speaker states, including both speaker transitions and silent moments. In the four-person setup, the model achieves an accuracy of 48%, while in the three-person setup, it reaches 55%. It is important to note that our model is also capable of predicting silent segments. As a result, the baseline accuracy for random guessing is 20% (for four speakers plus silence) and 25% (for three speakers plus silence), respectively.

7.1.2 Gaze Prediction. Table 2 presents quantitative evaluation of gaze prediction models, demonstrating that pre-trained models consistently outperform non-pre-trained baselines. Our SAT5 model achieves the highest precision, significantly surpassing standard transformer-based models such as BERT and T5. This improvement highlights the effectiveness of speaker-aware embeddings,

Table 1: Quantitative results of speaker state modeling, averaged over 10-fold cross-validation on the test sets with varying participants.

Participants	Accuracy	F1-Score
3	0.5513	0.5553
4	0.4827	0.4495

which allow the model to capture conversational dynamics and gaze behavior more accurately than text-only approaches.

To further analyze the contributions of individual components in SAT5, we conduct an ablation study, evaluating two key variations: SAT5 w/o Speaker State Modeling (SSM) and SAT5 w/o Span Corruption (SP). As shown in Table 2, both variations exhibit noticeable drops in performance, confirming the necessity of these multi-task learning objectives. The degradation in SAT5 w/o SSM suggests that explicitly modeling speaker transitions plays a critical role in gaze prediction, as it helps the model better understand when and where participants shift their attention. Similarly, the performance drop in SAT5 w/o SP underscores the importance of span corruption pretraining, which enhances the model's ability to handle pauses, interruptions, and incomplete utterances — common challenges in real-world multi-party conversations.

Additionally, we examine the impact of chunk size on gaze prediction accuracy to determine the optimal temporal resolution for capturing gaze shifts. As presented in Table 3, we experimented with different chunk sizes, including 1, 5, and 9 frames, to assess their effect on contextual modeling. The results indicate that a chunk size of 5 frames achieves the best balance between capturing contextual dependencies and mitigating noise from erratic gaze shifts. Smaller chunk sizes (e.g., 1 frame) result in excessive sensitivity to short-term variations, reducing robustness, while larger chunk sizes (e.g., 9 frames) lead to excessive smoothing, potentially overlooking finer temporal variations in gaze behavior. This analysis confirms that a 5-frame chunk size provides the optimal trade-off, allowing the SAT5 model to maintain temporal coherence while adapting to the dynamics of multi-party interactions.

These findings collectively demonstrate the effectiveness of SAT5 in leveraging speaker-aware multimodal adaptation for gaze prediction. The integration of speaker representations, combined with multi-task learning and optimal temporal segmentation, enables the model to achieve state-of-the-art performance in predicting eye gaze in multi-party conversations.

Table 2: Quantitative results of gaze prediction, averaged over 10-fold cross-validation on the test sets.

	Precision	Recall	F1-Score
RNN	0.4609	0.3961	0.4132
GRU	0.4628	0.3958	0.4134
LSTM	0.4786	0.4029	0.4226
BERT	0.5556	0.4508	0.4831
T5	0.5601	0.4635	0.4954
SAT5	**0.6680**	**0.5746**	**0.6122**
SAT5 w/o SSM	0.6132	0.4673	0.5094
SAT5 w/o SP	0.6504	0.5242	0.5685

Table 3: Quantitative results of gaze prediction, averaged over 10-fold cross-validation on the test sets with varying chunk sizes.

Chunk Size	Precision	Recall	F1-Score
1	0.6508	0.5050	0.5515
5	**0.6680**	**0.5746**	**0.6122**
9	0.6384	0.5477	0.5833

8 DISCUSSION AND CONCLUSION

In this paper, we present SAT5, a novel speaker-aware adaptation of the T5 language model, designed to capture conversational dynamics in multi-party settings. Our key contributions include: (1) introducing a circular speaker embedding system that encodes speaker identities as angular coordinates on a unit circle; (2) designing a multi-task pretraining objective that jointly models speaker state prediction and span corruption; and (3) applying SAT5 to the downstream task of gaze prediction, demonstrating its ability to model the interplay between linguistic cues and speaker transitions.

Experimental results show that SAT5 achieves state-of-the-art performance on gaze prediction, significantly outperforming baseline models and highlighting the value of integrating speaker-awareness into pre-trained language models. In addition, our speaker embedding strategy is model-agnostic and can be incorporated into any encoder that accepts token-level inputs, such as RNNs or BERT. While this work focuses on adapting T5, evaluating the effectiveness of the embedding design across different backbone models remains an important direction for future research.

Despite these contributions, several limitations remain. First, although our InHouseDataset includes 21 participants, which is comparable to datasets used in prior studies of multiparty conversation, it is limited to three-party interactions and was collected in a controlled laboratory setting. We plan to expand the dataset to include four- and five-party conversations to improve generalizability. Second, SAT5 currently focuses on gaze and text modalities. Although vision-based features are not incorporated in this work, integrating visual cues such as facial expressions and gestures represents a promising direction for future research and may further enhance model performance. Lastly, although SAT5 achieves strong results, additional improvements and validation are needed for real-world deployment. In future work, we will investigate model optimization techniques, reduce inference latency, and evaluate the system in live, interactive environments.

9 SAFE AND RESPONSIBLE INNOVATION STATEMENT

Our research prioritizes ethical and responsible innovation by exclusively using anonymized and Institutional Review Board (IRB)-approved data from consenting adult participants. We emphasize data privacy through secure handling and storage protocols, and the models developed do not infer or reveal sensitive personal information. To mitigate potential bias, our dataset includes diverse speakers and multimodal cues, ensuring inclusivity and minimizing representational harm. While our model enhances gaze-aware interaction, we acknowledge the risk of misuse in surveillance or deceptive human-agent systems, and advocate for its applications in transparent, user-consensual contexts such as assistive technology or educational tools.

ACKNOWLEDGMENTS

This work is supported in part by US NSF IIS grant 2005430.

REFERENCES

[1] Reginald B Adams Jr and Robert E Kleck. 2005. Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion* 5, 1 (2005), 3.

[2] Raksha Agarwal and Niladri Chatterjee. 2021. LangResearchLab_NC at CMCL2021 Shared Task: Predicting Gaze Behaviour Using Linguistic Features and Tree Regressors. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (Eds.). Association for Computational Linguistics, Online, 79–84. doi:10.18653/v1/2021.cmcl-1.8

[3] Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. 2017. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics* 28, 3 (2017), 449–483.

[4] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. 2020. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5396–5406.

[5] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Yu Ding, Yuting Zhang, Meihua Xiao, and Zhigang Deng. 2017. A multifaceted study on eye contact based speaker identification in three-party conversations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3011–3021.

[7] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. 2007. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin* 133, 4 (2007), 694.

[8] Tzvi Ganel, Yonatan Goshen-Gottstein, and Melvyn A Goodale. 2005. Interactions between the processing of gaze direction and facial expression. *Vision research* 45, 9 (2005), 1191–1200.

[9] Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2041–2044.

[10] Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. MPC-BERT: A pre-trained language model for multi-party conversation understanding. *arXiv preprint arXiv:2106.01541* (2021).

[11] Hang Guo, Zhengxi Hu, and Jingtai Liu. 2022. Mgtr: End-to-end mutual gaze detection with transformer. In *Proceedings of the Asian Conference on Computer Vision*. 1590–1605.

[12] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764.

[13] Aobo Jin, Qixin Deng, and Zhigang Deng. 2022. S2M-Net: Speech Driven Three-party Conversational Motion Synthesis Networks. In *Proceedings of the 15th ACM SIGGRAPH Conference on Motion, Interaction and Games*. 2:1–2:10.

[14] Aobo Jin, Qixin Deng, Yuting Zhang, and Zhigang Deng. 2019. A deep learning-based model for head and eye motion generation in three-party conversations. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2, 2 (2019), 1–19.

[15] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–30.

[16] Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The AMI meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*. 1–4.

[17] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[18] Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who Is Speaking to Whom? Learning to Identify Utterance Addressee in Multi-Party Conversations. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 1909–1919. doi:10.18653/v1/D19-1199

[19] Meng-Chen Lee and Zhigang Deng. 2024. Online Multimodal End-of-Turn Prediction for Three-party Conversations. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 57–65.

[20] Meng-Chen Lee, Wu Angela Li, and Zhigang Deng. 2024. A Computational Study on Sentence-based Next Speaker Prediction in Multiparty Conversations. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–4.

[21] Meng-Chen Lee, Mai Trinh, and Zhigang Deng. 2023. Multimodal Turn Analysis and Prediction for Multi-party Conversations. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 436–444.

[22] Bai Li and Frank Rudzicz. 2021. TorontoCL at CMCL 2021 Shared Task: RoBERTa with Multi-Stage Fine-Tuning for Eye-Tracking Prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (Eds.). Association for Computational Linguistics, Online, 85–89. doi:10.18653/v1/2021.cmcl-1.9

[23] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[24] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Alexander Koller, Gabriel Skantze, Filip Jurcicek, Masahiro Araki, and Carolyn Penstein Rose (Eds.). Association for Computational Linguistics, Prague, Czech Republic, 285–294. doi:10.18653/v1/W15-4640

[25] Byung-Doh Oh. 2021. Team Ohio State at CMCL 2021 Shared Task: Fine-Tuned RoBERTa for Eye-Tracking Data Prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (Eds.). Association for Computational Linguistics, Online, 97–101. doi:10.18653/v1/2021.cmcl-1.11

[26] Kazuhiro Otsuka. 2011. Multimodal conversation scene analysis for understanding people's communicative behaviors in face-to-face meetings. In *Proceedings of Symposium on Human Interface 2011*. Springer, 171–179.

[27] Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2133–2143.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[30] Magdalena Rychlowska, Leah Zinner, Serban C Musca, and Paula M Niedenthal. 2012. From the eye to the heart: eye contact triggers emotion simulation. In *Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction*. 1–7.

[31] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. 2018. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 8615–8621.

[32] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. 2022. End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2192–2200.

[33] Peter Vickers, Rosa Wainwright, Harish Tayyar Madabushi, and Aline Villavicencio. 2021. CogNLP-Sheffield at CMCL 2021 Shared Task: Blending Cognitively Inspired Features with Transformer-based Language Models for Predicting Eye Tracking Patterns. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus (Eds.). Association for Computational Linguistics, Online, 125–133. doi:10.18653/v1/2021.cmcl-1.16

[34] Binglu Wang, Chenxi Guo, Yang Jin, Haisheng Xia, and Nian Liu. 2024. TransGOP: transformer-based gaze object prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 10180–10188.

[35] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. 2024. LLM Knows Body Language, Too: Translating Speech Voices into Human Gestures. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5004–5013.