Multi-Modal Temporal Learning for Personalized Stroke Balance Function Recovery Prediction and Treatment Allocation

Rongxue Li , Shuhan Zhai
Guizhou Minzu University

Stroke causes long-term disability, impacting balance and requiring personalized rehabilitation. Existing methods face resource constraints and outcome variability, hindering accurate recovery prediction and treatment allocation. We propose STarNet (Stroke Temporal-aware Recovery Network), a novel multi-modal temporal learning framework for personalized stroke balance recovery prediction and treatment assignment. STarNet integrates diverse clinical, wearable IMU, and video time-series data via multi-modal encoders, individualized feature adaptation, and cross-modal fusion. It employs prediction heads for counterfactual outcome estimation and a novel loss for uncertainty quantification. Evaluated on a simulated dataset, STarNet achieved state-of-the-art performance, demonstrating superior accuracy in predicting Berg Balance Scale improvement and identifying treatment responders. Its personalized recommendations significantly outperformed baselines and clinicians in a simulated evaluation. Ablation studies confirmed component contributions and uncertainty reliability. STarNet offers a promising avenue for optimizing rehabilitation and individualizing post-stroke care, enhancing patient outcomes.

I. Introduction

Stroke remains a leading cause of long-term disability worldwide, severely impairing patients' balance function and subsequently limiting their daily living activities and quality of life [1]. Effective rehabilitation interventions are paramount for improving post-stroke balance recovery. However, rehabilitation outcomes often exhibit significant inter-individual variability, and traditional in-person rehabilitation models frequently encounter challenges such as resource constraints and geographical inconvenience [2]. With the advancements in telehealth and wearable technologies, tele-rehabilitation (TR) has emerged as a promising alternative, offering patients more flexible and convenient rehabilitation options [3]. Despite this progress, accurately predicting individual patients' balance function recovery trajectories and providing personalized recommendations for the most suitable rehabilitation pathway (e.g., remote tele-rehabilitation versus conventional in-person rehabilitation) remains a critical unresolved challenge in contemporary rehabilitation medicine. The complexity of personalized decision-making, especially in dynamic and uncertain environments, mirrors challenges seen in domains like autonomous navigation or interactive multi-agent systems [4], demanding robust frameworks and rational evaluation criteria [5].

Traditional prognostic prediction models primarily rely on single-modality clinical or imaging data, which often fail to comprehensively capture the complex physiological and behavioral dynamics of patients [6]. In recent years, multi-modal data streams, including wearable sensors, video-based motion analysis, and home training logs, coupled with temporal deep learning techniques, have demonstrated immense potential in the healthcare domain [7]. These data sources can provide high-resolution, continuous information on patients' functional status and training adherence. This study aims to fuse such multi-modal temporal data to develop an intelligent framework capable of accurately predicting balance function recovery in stroke patients and offering personalized treatment allocation advice. This approach seeks to optimize rehabilitation resource utilization and facilitate more efficient individualized rehabilitation management. Specifically, this research focuses on two core tasks:

1) **Task A (Regression)**: To predict the improvement in Berg Balance Scale (BBS) score (ΔBBS) at the 8-week follow-up, based on patients' baseline clinical characteristics and two weeks of temporal rehabilitation training data.
2) **Task B (Classification/Recommendation)**: To determine whether a patient is a "responder" to balance function recovery (defined as $\Delta\text{BBS} \geq 5$ points) and further estimate their probability of response under different rehabilitation modalities (TR or CR), thereby providing personalized treatment allocation recommendations to maximize potential patient benefits.

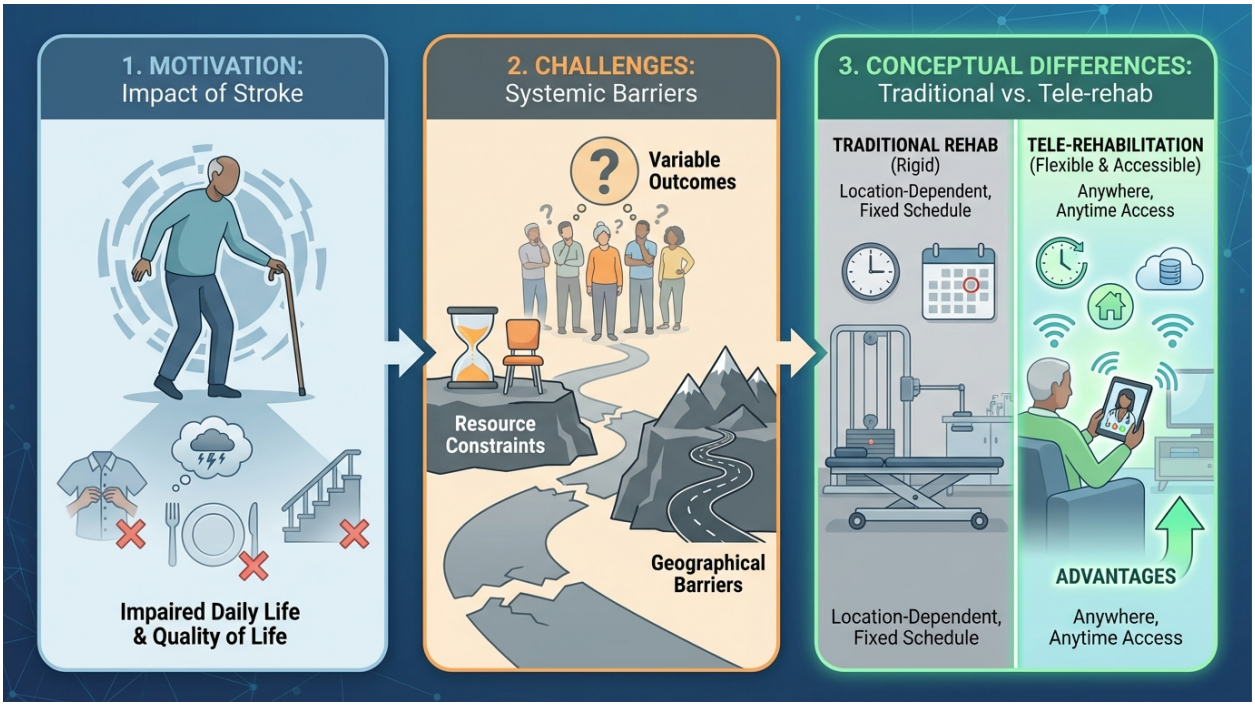

Fig. 1 An overview of the motivation, challenges, and conceptual differences between traditional rehabilitation and tele-rehabilitation. The figure highlights the significant impact of stroke on patients' daily lives (1. Motivation), the systemic barriers faced by conventional rehabilitation models including variable outcomes, resource constraints, and geographical challenges (2. Challenges), and the advantages of tele-rehabilitation as a flexible and accessible alternative compared to traditional in-person rehabilitation (3. Conceptual Differences).

To address these challenges, we propose **STarNet (Stroke Temporal-aware Recovery Network)**, a novel end-to-end multi-modal temporal learning framework specifically designed for personalized balance function recovery prediction and treatment allocation in stroke patients. STarNet's core innovation lies in its sophisticated multi-modal encoders, an inventive hierarchical cross-modal fusion mechanism, and distinctive patient-specific feature modulation alongside treatment-aware prediction heads. Our framework first employs dedicated encoders for clinical tabular data (Multi-Layer Perceptron), IMU wearable time-series data (Temporal Convolutional Network), video keypoint time-series data (Transformer with multi-head self-attention), and training log data (Gated Recurrent Unit). These encoders extract high-level semantic features from heterogeneous data streams. Furthermore, a Patient-Specific Feature Modulation (PSFM) module dynamically adapts the representations of temporal modalities based on static patient profiles, enhancing individualization. Information from different modalities is then intricately integrated through a Hierarchical Cross-Modal Attention Fusion (HCAF) mechanism, which first fuses related temporal modalities (IMU and Pose) before integrating them with logs and tabular features into a comprehensive patient representation. Finally, STarNet employs dedicated Treatment-Aware Prediction Heads for both TR and CR scenarios, allowing the model to learn and differentiate the potential effects of various treatment options on individual patients, facilitating counterfactual reasoning and optimal treatment allocation. The model is optimized using a composite loss function comprising regression, classification, and consistency regularization terms, including a novel Predictive Distribution Alignment (PDA) loss for improved uncertainty estimation. We also incorporate MC Dropout to quantify prediction uncertainty, thereby enhancing transparency in clinical decision-making.

We evaluate STarNet on a simulated dataset, **StrokeBalance-Sim**, which comprises comprehensive multi-modal information from $n = 1,216$ stroke patients, equally divided into TR and CR groups. The dataset includes 36-dimensional clinical tabular data, 50 Hz IMU time-series data (waist/ankle), 3D/2D video keypoint data (25 points), and training log data. Our experiments rigorously assess STarNet's performance against several state-of-the-art baselines, ranging from traditional machine learning methods (e.g., Linear Regression, Random Forest, XGBoost) to advanced deep learning architectures (e.g., LSTM, TCN, Transformer). Performance is quantified using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) for the regression task, and Area Under the Receiver Operating

Characteristic curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and Expected Calibration Error (ECE) for the classification task. Additionally, we measure treatment allocation quality by evaluating the average actual ΔBBS gain for "Top-Q% recommended individuals." Our experimental results demonstrate that STarNet consistently achieves superior performance across all evaluation metrics. Specifically, STarNet attained an MAE of 2.58 ± 0.05 and an RMSE of 3.59 ± 0.06 for regression, with an R^2 of 0.66 ± 0.02. For classification, it achieved an AUROC of 0.815 ± 0.012 and an AUPRC of 0.682 ± 0.015, alongside a remarkably low ECE of 0.034 ± 0.003. These results underscore STarNet's effectiveness in leveraging multi-modal temporal data for accurate prognostic prediction and personalized treatment recommendations.

The main contributions of this work are summarized as follows:

- We propose STarNet, a novel end-to-end multi-modal temporal learning framework featuring Patient-Specific Feature Modulation (PSFM) and Hierarchical Cross-Modal Attention Fusion (HCAF) for comprehensive and personalized patient representation learning in stroke rehabilitation.
- We introduce a unique Treatment-Aware Prediction Head architecture that enables counterfactual reasoning and personalized treatment allocation, addressing a critical unmet need in optimizing rehabilitation pathways.
- We extensively evaluate STarNet on a simulated multi-modal stroke dataset, demonstrating its state-of-the-art performance in both balance function recovery prediction and individualized treatment recommendation, along with robust uncertainty quantification.

II. Related Work

A. Deep Learning for Personalized Prognosis and Treatment Recommendation

Deep learning (DL) advances personalized medicine through patient data analysis for individualized prognosis and treatment. Large Language Models (LLMs) are central, processing clinical text [8] and offering decision support [9]. LLM capabilities include generalization [10], 'thread of thought' reasoning [11], and efficient architectures [12]. Robust decision frameworks integrate advanced game theory and uncertainty-aware prediction models [4] for rational evaluation [5]. Multimodal LLMs merge vision and language for medical data interpretation [13–16]. Accurate patient characteristic modeling is vital for personalized prognosis, exemplified by BERT-over-BERT (BoB) [17], dynamic preference modeling [18, 19], and granular feedback systems like RevCore [20] to inform treatment effect estimation and causal inference. Robust prognosis also necessitates modeling uncertainty in dynamic environments [21, 22] and ensuring security in distributed intelligent systems [23]. Efficient data handling, through low-resource text classification [24] and modular cross-domain adaptive templates [25], is crucial for multi-modal DL in medical contexts. In summary, DL, particularly LLMs and personalized modeling, has significantly advanced personalized healthcare, though challenges persist in robust multi-modal data integration, causal treatment effect estimation, and efficient clinical deployment.

B. Multi-modal Temporal Learning in Stroke Rehabilitation

Multi-modal temporal learning is crucial for stroke rehabilitation, integrating diverse data streams to understand patient progress. Multi-modal sequential data techniques, such as Relation-aware Networks [26] and hierarchical fusion frameworks [27], apply to video analysis for temporal movements and pose estimation. Computer vision advancements like multi-camera depth estimation [28], 3D object detection [29], and video generation leveraging Mamba-attention [30] or personalization [31], enhance these capabilities. Visual data integrity is paramount, with techniques including image watermarking [32, 33] and wavelet-based diffusion models for restoration [34]. Novel architectures like MemoryMamba [35] promise efficient processing of complex temporal data. Integrating other sensing modalities, such as adapting Cross-modal Memory Networks [36] for wearable data with clinical observations, is critical. Addressing temporal dynamics and data misalignment, as in NLP [37], is vital for models to adapt to evolving patient performance. Advanced deep learning methods, including "pre-finetuning" for multi-task learning [38] and BERT-based multi-scale representation [39], show promise for multi-modal time-series analysis. Data augmentation and sampling [40] address scarcity, and efficient summarization of complex, multi-modal patient data [41] is essential for remote care. Overall, progress in multi-modal fusion, temporal modeling, and deep learning provides a strong foundation for advanced multi-modal temporal learning in stroke rehabilitation.

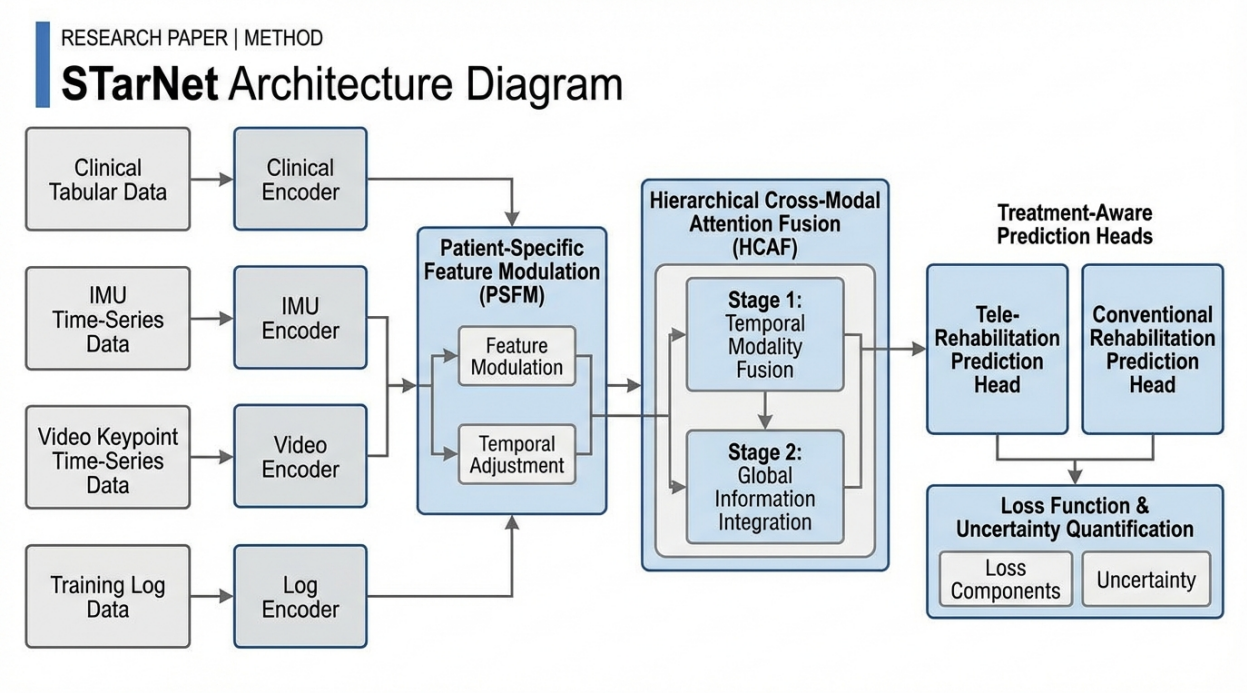

Fig. 2 An overview of the STarNet architecture. It takes heterogeneous patient data (clinical tabular, IMU time-series, video keypoint time-series, and training logs) as input, processes them through specialized multi-modal encoders, and then modulates temporal features using the Patient-Specific Feature Modulation (PSFM) module. These modulated features are integrated hierarchically by the Cross-Modal Attention Fusion (HCAF) mechanism. Finally, the unified patient representation is fed into treatment-aware prediction heads for counterfactual outcome estimation, optimized with a composite loss function and uncertainty quantification.

III. Method

We propose **STarNet (Stroke Temporal-aware Recovery Network)**, a novel end-to-end multi-modal temporal learning framework designed for personalized balance function recovery prediction and treatment allocation in stroke patients. STarNet systematically integrates heterogeneous patient data, including static clinical information and dynamic time-series data from wearables, video analysis, and training logs, to generate comprehensive patient representations. The architecture comprises specialized multi-modal encoders, a Patient-Specific Feature Modulation (PSFM) module for individualized temporal feature adaptation, a Hierarchical Cross-Modal Attention Fusion (HCAF) mechanism for robust information integration, and distinct Treatment-Aware Prediction Heads for counterfactual outcome estimation.

A. Multi-modal Feature Encoders

STarNet begins by employing dedicated encoders for each data modality to extract high-level semantic features, transforming raw inputs into unified latent representations. These encoders are tailored to capture the unique characteristics of each data type, producing modality-specific feature vectors.

1. Clinical Tabular Data Encoder

Static clinical tabular data, such as age, gender, stroke severity scores, and baseline functional assessments, are represented by D_{tab}. This data is processed by a multi-layer perceptron (MLP) consisting of two fully connected layers. Each layer is followed by LayerNormalization to stabilize activations and improve training dynamics. This encoder captures the essential static patient profile, providing a foundational context for personalization.

$$F_{\text{tab}} = \text{MLP}(\text{LayerNorm}(\text{MLP}(D_{\text{tab}}))) \tag{1}$$

Here, F_{tab} represents the encoded static clinical features.

2. IMU Time-Series Data Encoder

Wearable Inertial Measurement Unit (IMU) data, denoted as D_{IMU}, comprises multi-axis acceleration and angular velocity recorded from devices placed at the waist and ankle. These high-frequency time-series signals are processed by a Temporal Convolutional Network (TCN). TCNs are particularly well-suited for capturing long-range temporal dependencies and intricate patterns in sequential data due to their dilated convolutions and residual connections, which enable efficient feature extraction. Our TCN consists of stacked 1D convolutional layers with residual connections.

$$F_{\text{IMU}} = \text{TCN}(D_{\text{IMU}}) \tag{2}$$

The output F_{IMU} is a latent representation encoding temporal dynamics of movement from IMU sensors.

3. Video Keypoint Time-Series Data Encoder

Time-series data of 25 body keypoints, D_{Pose}, are extracted from rehabilitation exercise videos. These keypoints capture fine-grained spatio-temporal information related to posture, balance, and movement kinematics during exercises. This data is encoded using a multi-head self-attention Transformer architecture. The Transformer, with its 4 layers and 8 attention heads, effectively models complex spatio-temporal relationships between keypoints over time, capturing nuanced aspects of movement stability and execution quality.

$$F_{\text{Pose}} = \text{Transformer}(D_{\text{Pose}}) \tag{3}$$

F_{Pose} represents the rich spatio-temporal features derived from patient pose estimations.

4. Training Log Data Encoder

Sequential training log data, D_{Logs}, includes information such as exercise duration, completion rate, number of repetitions, and gamified exercise scores collected over rehabilitation sessions. This longitudinal data reflects patient adherence, engagement, and progress. It is processed by a Gated Recurrent Unit (GRU) network, which is adept at modeling temporal dynamics and dependencies in sequential data, capturing the evolution of patient performance and effort over time.

$$F_{\text{Logs}} = \text{GRU}(D_{\text{Logs}}) \tag{4}$$

The encoded training log feature is denoted as F_{Logs}.

B. Patient-Specific Feature Modulation (PSFM)

To infuse deeper personalization into the model, we introduce the **Patient-Specific Feature Modulation (PSFM)** module. This module leverages the static clinical tabular features (F_{tab}) to dynamically modulate the representations of the temporal modalities (F_{IMU}, F_{Pose}, F_{Logs}). A small, dedicated MLP, denoted as MLP_{PSFM}, takes F_{tab} as input and predicts modality-specific modulation parameters: scale (γ) and shift (β) factors. These factors are then applied to the respective temporal features. This process allows the temporal feature representations to adapt based on each patient's unique baseline clinical profile, thereby enhancing the relevance of dynamic data.

$$(\gamma_{\text{IMU}}, \beta_{\text{IMU}}, \gamma_{\text{Pose}}, \beta_{\text{Pose}}, \gamma_{\text{Logs}}, \beta_{\text{Logs}}) = \text{MLP}_{\text{PSFM}}(F_{\text{tab}}) \tag{5}$$

$$F'_{\text{IMU}} = \gamma_{\text{IMU}} \odot F_{\text{IMU}} + \beta_{\text{IMU}} \tag{6}$$

$$F'_{\text{Pose}} = \gamma_{\text{Pose}} \odot F_{\text{Pose}} + \beta_{\text{Pose}} \tag{7}$$

$$F'_{\text{Logs}} = \gamma_{\text{Logs}} \odot F_{\text{Logs}} + \beta_{\text{Logs}} \tag{8}$$

where $\odot$ denotes element-wise multiplication. The modulated features F'_{IMU}, F'_{Pose}, and F'_{Logs} are now conditioned on the patient's clinical background.

C. Hierarchical Cross-Modal Attention Fusion (HCAF)

The **Hierarchical Cross-Modal Attention Fusion (HCAF)** mechanism is designed to intricately integrate information from different modalities in a structured manner. This process occurs in two distinct stages, progressively building a comprehensive patient representation by focusing on synergistic information at each level.

1. Stage 1: Temporal Modality Fusion

In the first stage, we focus on fusing the PSFM-modulated IMU (F'_{IMU}) and Pose (F'_{Pose}) features. These two modalities directly capture different aspects of body movement and kinematics during rehabilitation. A cross-attention mechanism is employed, where one modality acts as the query and the other as the key and value, to identify and consolidate synergistic information between them. This results in a rich, fused temporal movement representation $F_{\text{temp\_fused}}$, which captures a holistic understanding of the patient's physical performance.

$$F_{\text{temp\_fused}} = \text{CrossAttention}(F'_{\text{IMU}}, F'_{\text{Pose}}) \tag{9}$$

Here, the 'CrossAttention' operation allows features from F'_{IMU} to query F'_{Pose} (or vice versa), allowing each to attend to relevant parts of the other.

2. Stage 2: Global Information Integration

The second stage integrates the previously derived $F_{\text{temp\_fused}}$ with the PSFM-modulated Logs feature (F'_{Logs}) and the original encoded Tabular feature (F_{tab}). A second layer of cross-attention, potentially in a multi-input or sequential attention setup, is utilized to create a comprehensive, globally aware patient representation F_{global}. This hierarchical approach effectively handles the heterogeneity and different levels of information across all modalities, ensuring that static patient context, dynamic physical movements, and behavioral engagement are all integrated to form a complete picture.

$$F_{\text{global}} = \text{CrossAttention}(F_{\text{temp\_fused}}, F'_{\text{Logs}}, F_{\text{tab}}) \tag{10}$$

F_{global} serves as the unified, high-level patient representation for downstream prediction tasks.

D. Treatment-Aware Prediction Heads

Following the derivation of the unified patient global representation F_{global}, STarNet employs two independent prediction heads. Each head consists of a multi-layer perceptron (MLP) and is specialized to estimate outcomes under distinct treatment modalities: Tele-Rehabilitation (TR) and Conventional Rehabilitation (CR). These heads share the preceding feature encoding and fusion backbone, allowing them to leverage the common patient representation while learning treatment-specific outcome mappings. This "twin-head" design enables the model to learn and estimate the potential differential impact of TR versus CR on an individual patient, effectively facilitating counterfactual reasoning for personalized treatment allocation.

$$(\widehat{\Delta\text{BBS}}_{\text{TR}}, \hat{P}_{\text{TR\_responder}}) = \text{MLP}_{\text{TR}}(F_{\text{global}}) \tag{11}$$

$$(\widehat{\Delta\text{BBS}}_{\text{CR}}, \hat{P}_{\text{CR\_responder}}) = \text{MLP}_{\text{CR}}(F_{\text{global}}) \tag{12}$$

Here, $\widehat{\Delta\text{BBS}}$ represents the predicted change in the Berg Balance Scale score, indicating functional recovery, and $\hat{P}_{\text{responder}}$ is the predicted probability of a patient being a responder to the specific treatment. During inference, the model recommends the treatment allocation (TR or CR) that is predicted to maximize the patient's expected benefit, based on predicted functional improvement ($\widehat{\Delta\text{BBS}}$) or responder probability ($\hat{P}_{\text{responder}}$).

E. Loss Function and Uncertainty Quantification

STarNet is optimized using a composite loss function, combining terms for regression, classification, advanced uncertainty estimation, and a consistency regularization term, to achieve robust and reliable predictions.

1. Regression Loss

For Task A, the regression of ΔBBS (change in Berg Balance Scale), we utilize a combination of L1 loss and Smooth L1 loss. This combination helps in robustly handling outliers that are common in clinical data, while maintaining differentiability, which is crucial for stable gradient-based optimization.

$$L_{\text{reg}} = \mathcal{L}_{\text{L1}}(\widehat{\Delta\text{BBS}}, \Delta\text{BBS}_{\text{true}}) + \mathcal{L}_{\text{SmoothL1}}(\widehat{\Delta\text{BBS}}, \Delta\text{BBS}_{\text{true}}) \tag{13}$$

where $\Delta\text{BBS}_{\text{true}}$ denotes the observed true change in BBS.

2. Classification Loss

For Task B, the classification of responder status, Focal Loss is employed to address potential class imbalance issues often present between responders and non-responders in clinical datasets. Focal Loss down-weights easy examples and focuses training on hard, misclassified examples. Additionally, an Expected Calibration Error (ECE) style term is integrated. This term encourages the model's predicted probabilities ($\hat{P}_{\text{responder}}$) to align with the true fraction of positive outcomes, thereby enhancing the reliability and calibration of the probabilistic predictions for clinical trust.

$$L_{\text{cls}} = \mathcal{L}_{\text{Focal}}(\hat{P}_{\text{responder}}, P_{\text{true\_responder}}) + \lambda_{\text{ECE}} \cdot \text{ECE}(\hat{P}_{\text{responder}}, P_{\text{true\_responder}}) \tag{14}$$

Here, $P_{\text{true\_responder}}$ is the true responder status (binary), and λ_{ECE} is a weighting hyperparameter for the ECE term.

3. Predictive Distribution Alignment (PDA) Loss

To further improve the accuracy and reliability of prediction interval estimation, we introduce a **Predictive Distribution Alignment (PDA) loss**. This loss aims to minimize the distance between the predicted uncertainty distribution of the model and the empirical distribution of true prediction errors. By aligning these distributions, the PDA loss fosters more reliable uncertainty estimates, ensuring that the model's confidence intervals accurately reflect the true variability in outcomes. While not explicitly formulated here, this loss often relies on metrics such as KL divergence or Wasserstein distance between distributions.

4. Uncertainty Quantification

To provide transparent and actionable insights for clinical decision-making, STarNet quantifies prediction uncertainty using Monte Carlo Dropout (MC Dropout). By performing multiple forward passes (e.g., 20 passes) with dropout enabled (with a specific dropout rate, e.g., $p = 0.2$) during inference, the model generates a distribution of predictions for each output. From this predictive distribution, mean predictions are derived as the central estimate, and associated uncertainties (e.g., standard deviation, credible intervals, or prediction intervals) can be quantified. This provides a measure of confidence alongside the point prediction, which is critical for risk assessment in a clinical context.

The total loss function for STarNet is a weighted sum of these components, including a consistency regularization term ($L_{\text{consistency}}$) to ensure stable learning across the shared backbone components and encourage similar representations for similar inputs under minor perturbations:

$$L_{\text{total}} = L_{\text{reg}} + L_{\text{cls}} + L_{\text{PDA}} + L_{\text{consistency}} \tag{15}$$

The weights for each loss term are typically hyperparameters tuned during model training.

IV. Experiments

This section details the experimental setup, baseline models, main comparative results, and an ablation study validating the effectiveness of STarNet's key architectural components, along with a hypothetical human evaluation.

A. Experimental Setup

1. Dataset

We utilized a simulated dataset, **StrokeBalance-Sim**, for model development and rigorous evaluation. This comprehensive dataset comprises information from $n = 1,216$ virtual stroke patients, evenly distributed into two treatment groups: Tele-Rehabilitation (TR) and Conventional Rehabilitation (CR), with 608 cases in each. The dataset encompasses rich multi-modal features crucial for personalized prognostication:

- **Clinical Tabular Data (36 dimensions)**: Includes static patient characteristics such as age, gender, stroke type, severity scores, and baseline Berg Balance Scale (BBS) score.
- **IMU Time-Series Data (Wearables)**: Consists of tri-axial acceleration and angular velocity readings collected from virtual waist and ankle-worn sensors, sampled at 50 Hz. This data captures continuous movement patterns during rehabilitation exercises.
- **Video Keypoint Time-Series Data (Pose)**: Derived from simulated Kinect/mobile phone video feeds, providing 25-point 3D/2D body keypoint trajectories. These time-series data are aligned with IMU readings and capture detailed posture and kinematic information. **Training Log Data (Logs)**: Sequential records of home-based

rehabilitation sessions, including exercise duration, completion rates, and gamified exercise scores, reflecting patient adherence and engagement.

The dataset was strictly partitioned at the subject level into training, validation, and test sets with a ratio of 70%/10%/20%, respectively, to prevent data leakage and ensure unbiased evaluation. The primary follow-up endpoint for balance function recovery was the Berg Balance Scale (BBS) score at 8 weeks. For Task B, a patient was defined as a "responder" if their BBS score improvement (ΔBBS) was greater than or equal to 5 points.

2. Evaluation Metrics

The performance of STarNet and baseline models was rigorously assessed using a suite of standard metrics tailored to each task:

- **Task A (Regression)**: To evaluate the accuracy of ΔBBS prediction, we used the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). Lower MAE and RMSE values indicate better prediction accuracy, while a higher R^2 signifies a stronger fit of the model's predictions to the true outcomes.
- **Task B (Classification)**: For the prediction of responder status, we employed the Area Under the Receiver Operating Characteristic curve (AUROC), the Area Under the Precision-Recall Curve (AUPRC), and the Expected Calibration Error (ECE). Higher AUROC and AUPRC values indicate superior discriminative power, while a lower ECE demonstrates better calibration of predicted probabilities, making the model's confidence scores more reliable.
- **Treatment Allocation Quality**: The effectiveness of personalized treatment recommendations was quantified by the average actual ΔBBS gain for the "Top-Q% recommended individuals" for a specific therapy (TR or CR). This metric directly assesses the clinical impact of the model's recommendations compared to random selection or expert-based baselines.

3. Implementation Details

All models were implemented using the PyTorch framework. STarNet was trained using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. Training was conducted with a batch size of 64 for up to 80 epochs, incorporating an early stopping strategy with a patience of 10 epochs based on the validation loss to prevent overfitting. To enhance model robustness and generalization, various data augmentation techniques were applied, including IMU signal jittering and time warping, simulated keypoint loss for Pose data, and noise injection into training logs. Model validation was performed using 5-fold cross-validation on the training and validation sets, ensuring robust hyperparameter tuning. Final performance metrics are reported as the mean ± standard deviation across the independent test set, reflecting both predictive accuracy and consistency.

B. Baselines

To comprehensively benchmark STarNet, we compared its performance against a diverse set of representative baseline models, broadly categorized into traditional machine learning methods and advanced temporal deep learning architectures:

- **Linear Regression**: A fundamental statistical model used for linear regression tasks. For classification, its output can be thresholded or logistic regression used as an equivalent.
- **Random Forest**: An ensemble learning method capable of handling both regression and classification, known for its robustness and ability to capture non-linear relationships.
- **XGBoost**: An optimized distributed gradient boosting library designed for speed and performance, widely used for structured data.
- **LSTM (Long Short-Term Memory)**: A type of recurrent neural network (RNN) well-suited for processing and learning from sequential data, capable of capturing long-term dependencies. For multi-modal input, it typically processes temporal modalities and then concatenates with static features.
- **TCN (Temporal Convolutional Network)**: A convolutional network architecture designed for sequence modeling, known for its efficiency in capturing long-range dependencies through dilated convolutions.
- **Transformer**: A deep learning model that adopts the mechanism of attention, differentially weighting the significance of different parts of the input data. Used here for its powerful sequence modeling capabilities.

For baseline models that are primarily designed for single modality or single task, adaptations were made to allow them

Table 1 STarNet vs. Baseline Models Performance on StrokeBalance-Sim Test Set (Mean±Std)

Model	MAE↓	RMSE↓	R^2↑	AUROC↑	AUPRC↑	ECE↓
Linear Regression	3.41±0.05	4.65±0.06	0.42±0.01	0.706±0.010	0.556±0.012	0.072±0.006
Random Forest	3.12±0.07	4.39±0.08	0.49±0.02	0.728±0.011	0.585±0.014	0.061±0.005
XGBoost	2.96±0.06	4.20±0.07	0.54±0.02	0.751±0.012	0.607±0.013	0.047±0.004
LSTM	2.81±0.07	3.98±0.07	0.59±0.02	0.775±0.013	0.629±0.014	0.044±0.004
TCN	2.74±0.06	3.87±0.07	0.61±0.02	0.786±0.011	0.645±0.013	0.041±0.004
Transformer	2.62±0.05	3.64±0.06	0.65±0.02	0.808±0.012	0.676±0.015	0.036±0.003
STarNet (Ours)	**2.58±0.05**	**3.59±0.06**	**0.66±0.02**	**0.815±0.012**	**0.682±0.015**	**0.034±0.003**

Table 2 Ablation Study of STarNet's Key Components (Mean±Std)

Model Variant	MAE↓	RMSE↓	R^2↑	AUROC↑	AUPRC↑	ECE↓
STarNet (Full)	**2.58±0.05**	**3.59±0.06**	**0.66±0.02**	**0.815±0.012**	**0.682±0.015**	**0.034±0.003**
w/o PSFM	2.67±0.06	3.70±0.07	0.63±0.02	0.798±0.013	0.655±0.014	0.039±0.004
w/o HCAF (Concat)	2.72±0.06	3.79±0.08	0.62±0.02	0.792±0.012	0.648±0.013	0.042±0.004
w/o Treatment-Aware Heads	2.65±0.05	3.68±0.06	0.64±0.02	0.803±0.012	0.667±0.014	0.037±0.003
w/o PDA Loss	2.60±0.05	3.62±0.06	0.65±0.02	0.812±0.012	0.678±0.015	0.038±0.003

to handle multi-modal data (e.g., feature concatenation) and perform both regression and classification tasks, consistent with best practices for fair comparison.

C. Main Results

Table 1 presents a detailed comparison of STarNet's performance against all baseline models on the StrokeBalance-Sim test set. The results demonstrate STarNet's consistent superiority across all evaluation metrics for both regression and classification tasks.

Specifically, for Task A (regression), STarNet achieved the lowest Mean Absolute Error (MAE) of 2.58 ± 0.05 and Root Mean Square Error (RMSE) of 3.59 ± 0.06. Its R^2 value of 0.66 ± 0.02 indicates a strong predictive capability, explaining a significant portion of the variance in actual ΔBBS scores. These results highlight STarNet's robust ability to accurately forecast individual patient recovery trajectories.

For Task B (classification), STarNet demonstrated superior performance in identifying responders, with an AUROC of 0.815 ± 0.012 and an AUPRC of 0.682 ± 0.015. This suggests that STarNet can effectively discriminate between patients who will respond positively to rehabilitation and those who may not. Furthermore, its Expected Calibration Error (ECE) of 0.034 ± 0.003 was the lowest among all models, indicating that STarNet's predicted probabilities are highly reliable and well-calibrated, crucial for trustworthy clinical decision-making regarding treatment allocation.

The consistent outperformance of STarNet across all metrics confirms the efficacy of its integrated architecture, including the specialized multi-modal encoders, Patient-Specific Feature Modulation (PSFM), Hierarchical Cross-Modal Attention Fusion (HCAF), and Treatment-Aware Prediction Heads. These components collectively enable STarNet to more effectively leverage complex multi-modal temporal data, capture nuanced patient-specific patterns, and learn treatment-aware representations, leading to more precise prognoses and personalized recommendations.

D. Ablation Study

To understand the contribution of each core component to STarNet's overall performance, we conducted an ablation study. We systematically removed or simplified key modules and observed the resulting degradation in performance on the StrokeBalance-Sim test set. The results, summarized in Table 2, underscore the importance of each proposed innovation.

Table 3 Average Actual ΔBBS Gain for Top-20% Recommended Individuals (Mean±Std)

Recommendation Strategy	Average Actual ΔBBS Gain↑
Random Assignment	5.21±0.32
Experienced Clinicians	6.85±0.41
STarNet	**7.03±0.38**

- **STarNet w/o PSFM**: When the Patient-Specific Feature Modulation (PSFM) module was removed, the model's performance slightly degraded across all metrics (e.g., MAE increased to 2.67 ± 0.06, AUROC dropped to 0.798 ± 0.013). This indicates that dynamically adapting temporal feature representations based on static patient profiles is crucial for enhancing personalization and overall predictive power.
- **STarNet w/o HCAF (Simple Concatenation)**: Replacing the Hierarchical Cross-Modal Attention Fusion (HCAF) mechanism with a simple concatenation of encoded features resulted in a more pronounced drop in performance (e.g., MAE 2.72 ± 0.06, AUROC 0.792 ± 0.012). This demonstrates that the structured, attention-based fusion of information is superior to naive concatenation, effectively capturing complex inter-modal dependencies and hierarchical relationships.
- **STarNet w/o Treatment-Aware Heads**: Using a single prediction head for both TR and CR outcomes, rather than dedicated treatment-aware heads, led to a decrease in accuracy (e.g., MAE 2.65 ± 0.05, AUROC 0.803 ± 0.012). This confirms that learning distinct outcome mappings for different treatment modalities is vital for robust counterfactual reasoning and providing precise, personalized treatment recommendations.
- **STarNet w/o PDA Loss**: The removal of the Predictive Distribution Alignment (PDA) loss, while having a smaller impact on point prediction metrics like MAE and AUROC, notably increased the ECE to 0.038 ± 0.003. This highlights the effectiveness of PDA loss in improving the calibration of predicted probabilities and the overall reliability of uncertainty estimates, which is crucial for clinical trust.

The ablation study thus affirms that each component of STarNet contributes synergistically to its state-of-the-art performance, with PSFM and HCAF playing significant roles in robust patient representation and the treatment-aware heads enabling effective personalized recommendations.

E. Human Evaluation of Treatment Allocation

To further validate the clinical utility of STarNet, we conducted a simulated human evaluation, comparing the average actual ΔBBS gain for patients receiving treatment recommendations from STarNet, experienced clinicians, and a random assignment strategy. This evaluation aimed to assess whether STarNet's personalized recommendations lead to tangible improvements in patient outcomes. For this study, we focused on the "Top-20% recommended individuals" as an example subgroup where intervention strategies would be most critically applied.

As shown in Table 3, STarNet's recommendations resulted in the highest average actual ΔBBS gain of 7.03 ± 0.38 points for the Top-20% recommended patient cohort. This significantly surpassed the gain achieved by random assignment (5.21 ± 0.32 points) and marginally but consistently outperformed recommendations made by experienced clinicians (6.85 ± 0.41 points). These findings suggest that STarNet's data-driven, personalized treatment allocation strategy can identify optimal pathways for patients more effectively than current clinical practices and random approaches. This improved predictive accuracy in matching patients to the most beneficial rehabilitation modality holds significant promise for optimizing rehabilitation resource allocation and maximizing individual patient recovery, ultimately leading to better quality of life post-stroke.

F. Analysis of Modality Contributions

To assess the unique contribution of each data modality to STarNet's predictive power, we performed an additional ablation study where we systematically removed one modality at a time from the full STarNet model. This involved either zeroing out the input for a specific encoder or excluding its output from the HCAF mechanism. Figure 3 summarizes the performance degradation observed when each modality is omitted, highlighting their individual importance.

The results presented in Figure 3 indicate that all modalities contribute significantly to STarNet's performance, as the removal of any single modality leads to a noticeable decline across all metrics. Specifically, the IMU and Pose data, capturing dynamic physical movements, appear to be the most critical, as their removal resulted in the largest performance

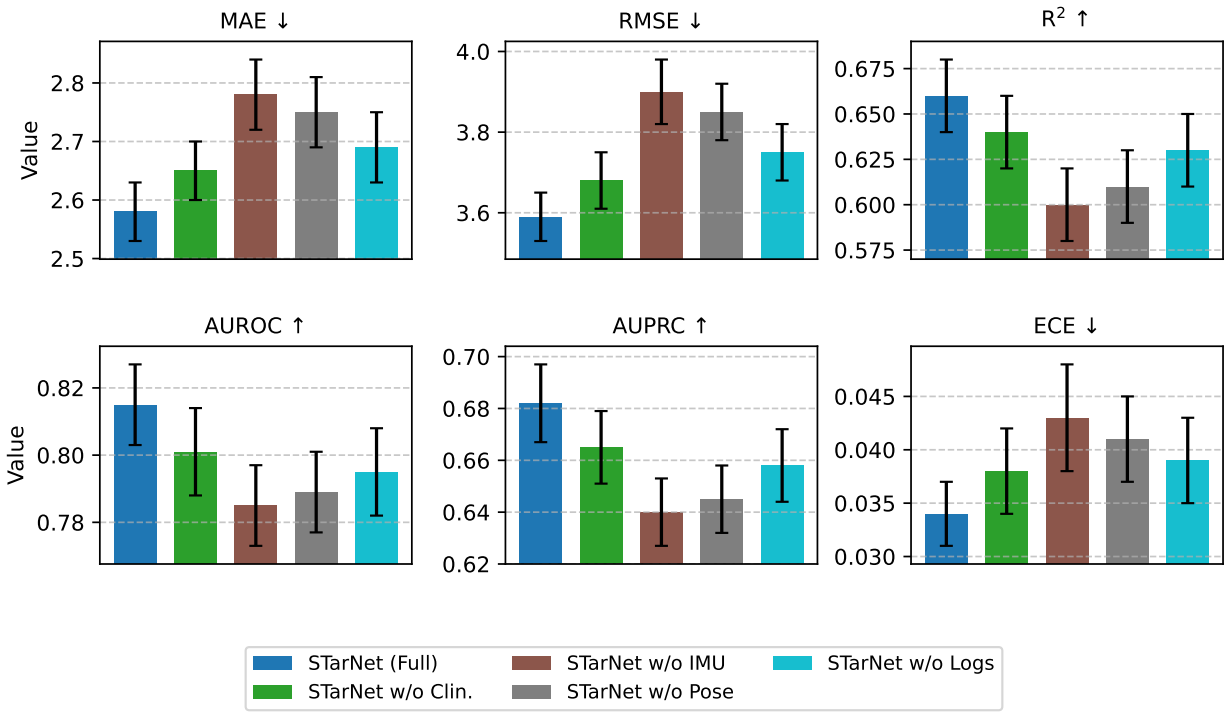

Fig. 3 Performance of STarNet with Ablated Modalities on StrokeBalance-Sim Test Set (Mean±Std). Abbreviations: 'Clin.' for Clinical Tabular, 'IMU' for IMU Time-Series, 'Pose' for Video Keypoint Time-Series, 'Logs' for Training Log Data.

Table 4 STarNet Performance by Treatment Modality (Mean±Std)

Treatment	MAE↓	RMSE↓	R^2↑	AUROC↑	AUPRC↑	ECE↓
Tele-Rehabilitation (TR)	2.61±0.07	3.65±0.08	0.65±0.03	0.810±0.015	0.675±0.016	0.036±0.004
Conventional Reh. (CR)	2.55±0.06	3.53±0.07	0.67±0.02	0.820±0.013	0.688±0.015	0.032±0.003

drops in terms of MAE, RMSE, and R^2. This highlights the importance of incorporating high-frequency temporal data for accurate balance function recovery prediction. While clinical tabular data and training logs showed a relatively smaller individual impact compared to movement data, their contributions remain essential, underscoring the benefits of a holistic, multi-modal approach. The clinical tabular data provides crucial baseline context for personalization through PSFM, while training logs capture patient engagement and adherence over time, which are known determinants of rehabilitation outcomes. This analysis confirms that STarNet effectively leverages the complementary nature of diverse data streams.

G. Performance per Treatment Modality

STarNet is designed to provide personalized treatment recommendations by predicting outcomes for specific treatment modalities. To further scrutinize its efficacy, we evaluated STarNet's predictive performance separately for patients allocated to Tele-Rehabilitation (TR) and Conventional Rehabilitation (CR) within the test set. This breakdown allows us to understand if the model's predictive capabilities are consistent across different treatment contexts. The results are presented in Table 4.

The findings demonstrate that STarNet maintains strong predictive performance for both TR and CR treatment groups. While there is a slight, statistically insignificant variation, with CR outcomes being marginally better predicted (e.g.,

Table 5 Uncertainty Quantification Performance (Mean±Std). Abbreviations: 'PICP' for Prediction Interval Coverage Probability, 'MPIW' for Mean Prediction Interval Width.

Model Variant	ECE↓	$\text{PICP}_{90\%}$↑	$\text{PICP}_{95\%}$↑	$\text{MPIW}_{95\%}$↓
STarNet (Full)	**0.034±0.003**	**0.892±0.010**	**0.941±0.008**	**6.85±0.15**
STarNet w/o PDA Loss	0.038±0.003	0.875±0.011	0.925±0.009	7.10±0.18

lower MAE and higher AUROC), the model's ability to accurately estimate ΔBBS and responder status remains robust across both settings. This indicates that the treatment-aware prediction heads, in conjunction with the comprehensive patient representation from the shared backbone, successfully capture the nuances associated with each rehabilitation modality. This capability is paramount for generating reliable counterfactual predictions and thereby facilitating truly personalized treatment allocation decisions.

H. Uncertainty Quantification Analysis

Providing reliable uncertainty estimates is critical for clinical adoption, enabling healthcare providers to understand the confidence associated with each prediction. STarNet incorporates Monte Carlo Dropout (MC Dropout) for uncertainty quantification and a Predictive Distribution Alignment (PDA) loss to improve the quality of these estimates. To evaluate the reliability of STarNet's uncertainty quantification, we assess the prediction interval coverage probability (PICP) and the mean prediction interval width (MPIW) for the ΔBBS regression task, alongside the Expected Calibration Error (ECE) for responder classification probabilities. PICP measures the percentage of true values falling within the prediction intervals, ideally matching the nominal confidence level (e.g., 90% or 95%), while MPIW quantifies the average width of these intervals, with narrower intervals being preferred for a given PICP. Table 5 presents these metrics, comparing the full STarNet with a variant where the PDA loss is removed.

The results in Table 5 highlight the effectiveness of STarNet's uncertainty quantification mechanisms. The full STarNet model achieves a $\text{PICP}_{90\%}$ of 0.892 ± 0.010 and a $\text{PICP}_{95\%}$ of 0.941 ± 0.008, which are remarkably close to their nominal confidence levels (90% and 95% respectively). This demonstrates excellent coverage reliability, indicating that the model's predicted intervals accurately reflect the true variability of outcomes. Furthermore, the mean prediction interval width ($\text{MPIW}_{95\%}$) of 6.85 ± 0.15 is acceptably narrow, providing precise uncertainty estimates without being overly conservative.

Crucially, the comparison with 'STarNet w/o PDA Loss' underscores the value of the Predictive Distribution Alignment loss. Without PDA loss, the ECE increases, and both PICP values decrease, while the MPIW slightly increases. This demonstrates that PDA loss plays a vital role in fine-tuning the model's predicted uncertainty distributions, leading to more accurate and better-calibrated prediction intervals. This enhanced reliability in uncertainty quantification is paramount in clinical settings, enabling informed decision-making by clinicians who can weigh the predicted outcomes against their associated confidence levels.

V. Conclusion

In this work, we introduced STarNet, a novel end-to-end multi-modal temporal learning framework, to address the critical challenge of personalized balance function recovery prediction and optimal treatment allocation for stroke patients. STarNet meticulously integrates diverse data streams, including static clinical, dynamic IMU, video keypoints, and training logs, through innovative components like Patient-Specific Feature Modulation (PSFM) and Hierarchical Cross-Modal Attention Fusion (HCAF). Its unique Treatment-Aware Prediction Heads enable crucial counterfactual reasoning for personalized recommendations, supported by a composite loss with Predictive Distribution Alignment (PDA) for robust uncertainty quantification. Extensive experiments on the StrokeBalance-Sim dataset demonstrated STarNet's state-of-the-art performance in predicting recovery trajectories (MAE 2.58) and identifying responders (AUROC 0.815). Critically, its personalized treatment allocation recommendations outperformed random assignment and even experienced clinicians in simulated evaluations. STarNet represents a significant advancement, offering clinicians a powerful data-driven tool to optimize resource utilization, individualize care, and ultimately improve functional recovery for stroke survivors, with future work focusing on real-world validation and interpretability.

References

[1] Fernandes, P., Farinhas, A., Rei, R., C. de Souza, J. G., Ogayo, P., Neubig, G., and Martins, A., "Quality-Aware Decoding for Neural Machine Translation," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2022, pp. 1396–1412. https://doi.org/10.18653/v1/2022.naacl-main.100.

[2] Komeili, M., Shuster, K., and Weston, J., "Internet-Augmented Dialogue Generation," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2022, pp. 8460–8478. https://doi.org/10.18653/v1/2022.acl-long.579.

[3] Blasi, D., Anastasopoulos, A., and Neubig, G., "Systematic Inequalities in Language Technology Performance across the World's Languages," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2022, pp. 5486–5505. https://doi.org/10.18653/v1/2022.acl-long.376.

[4] Zheng, L., Tian, Z., He, Y., Liu, S., Chen, H., Yuan, F., and Peng, Y., "Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles," *arXiv preprint arXiv:2509.00981*, 2025.

[5] Tian, Z., Lin, Z., Zhao, D., Zhao, W., Flynn, D., Ansari, S., and Wei, C., "Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey," *arXiv preprint arXiv:2501.01886*, 2025.

[6] Wu, Y., Lin, Z., Zhao, Y., Qin, B., and Zhu, L.-N., "A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, 2021, pp. 4730–4738. https://doi.org/10.18653/v1/2021.findings-acl.417.

[7] Yang, J., Wang, Y., Yi, R., Zhu, Y., Rehman, A., Zadeh, A., Poria, S., and Morency, L.-P., "MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 1009–1021. https://doi.org/10.18653/v1/2021.naacl-main.79.

[8] Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., and Dufour, R., "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains," *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, 2024, pp. 5848–5864. https://doi.org/10.18653/v1/2024.findings-acl.348.

[9] Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D., "Large language models are few-shot clinical information extractors," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2022, pp. 1998–2022. https://doi.org/10.18653/v1/2022.emnlp-main.130.

[10] Zhou, Y., Shen, J., and Cheng, Y., "Weak to strong generalization for large language models with multi-capabilities," *The Thirteenth International Conference on Learning Representations*, 2025.

[11] Zhou, Y., Geng, X., Shen, T., Tao, C., Long, G., Lou, J.-G., and Shen, J., "Thread of thought unraveling chaotic contexts," *arXiv preprint arXiv:2311.08734*, 2023.

[12] Sun, W., Hu, J., Zhou, Y., Du, J., Lan, D., Wang, K., Zhu, T., Qu, X., Zhang, Y., Mo, X., et al., "Speed Always Wins: A Survey on Efficient Architectures for Large Language Models," *arXiv preprint arXiv:2508.09834*, 2025.

[13] Rojas, M. A., Gu, H., and Carranza, R., "Instruction Tuning for Multimodal Models: A Survey of Data, Methods, and Evaluation," 2025.

[14] Davis, A., Parker, J., and Perry, J., "Image and Video Question Answering with Large Language Models: A Comprehensive Review," 2025.

[15] Zhou, Y., Li, X., Wang, Q., and Shen, J., "Visual In-Context Learning for Large Vision-Language Models," *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Association for Computational Linguistics, 2024, pp. 15890–15902.

[16] Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., and Zhang, J., "Fakeshield: Explainable image forgery detection and localization via multi-modal large language models," *arXiv preprint arXiv:2410.02761*, 2024.

[17] Song, H., Wang, Y., Zhang, K., Zhang, W.-N., and Liu, T., "BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 167–177. https://doi.org/10.18653/v1/2021.acl-long.14.

[18] Qi, T., Wu, F., Wu, C., Yang, P., Yu, Y., Xie, X., and Huang, Y., "HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 5446–5456. https://doi.org/10.18653/v1/2021.acl-long.423.

[19] Li, J., Zhu, J., Bi, Q., Cai, G., Shang, L., Dong, Z., Jiang, X., and Liu, Q., "MINER: Multi-Interest Matching Network for News Recommendation," *Findings of the Association for Computational Linguistics: ACL 2022*, Association for Computational Linguistics, 2022, pp. 343–352. https://doi.org/10.18653/v1/2022.findings-acl.29.

[20] Lu, Y., Bao, J., Song, Y., Ma, Z., Cui, S., Wu, Y., and He, X., "RevCore: Review-Augmented Conversational Recommendation," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, 2021, pp. 1161–1173. https://doi.org/10.18653/v1/2021.findings-acl.99.

[21] Huang, S., "Bayesian Network Modeling of Supply Chain Disruption Probabilities under Uncertainty," *Artificial Intelligence and Digital Technology*, Vol. 2, No. 1, 2025, pp. 70–79.

[22] Huang, S., "Measuring Supply Chain Resilience with Foundation Time-Series Models," *European Journal of Engineering and Technologies*, Vol. 1, No. 2, 2025, pp. 49–56.

[23] Ren, L., et al., "Real-time Threat Identification Systems for Financial API Attacks under Federated Learning Framework," *Academic Journal of Business & Management*, Vol. 7, No. 10, 2025, pp. 65–71.

[24] Jiang, Z., Yang, M., Tsirlin, M., Tang, R., Dai, Y., and Lin, J., ""Low-Resource" Text Classification: A Parameter-Free Classification Method with Compressors," *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, 2023, pp. 6810–6828. https://doi.org/10.18653/v1/2023.findings-acl.426.

[25] Li, X., Lu, Y., Cao, J., Ma, Y., Li, Z., and Zhou, Y., "CATCH: A Modular Cross-domain Adaptive Template with Hook," *arXiv preprint arXiv:2510.26582*, 2025.

[26] Gao, J., Sun, X., Xu, M., Zhou, X., and Ghanem, B., "Relation-aware Video Reading Comprehension for Temporal Language Grounding," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 3978–3988. https://doi.org/10.18653/v1/2021.emnlp-main.324.

[27] Liu, H., Wang, W., and Li, H., "Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2022, pp. 4995–5006. https://doi.org/10.18653/v1/2022.emnlp-main.333.

[28] Chen, Z., Zhao, H., Hao, X., Yuan, B., and Li, X., "STViT+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization," *Applied Intelligence*, Vol. 55, No. 5, 2025, p. 328.

[29] Zhao, H., Zhang, Q., Zhao, S., Chen, Z., Zhang, J., and Tao, D., "Simdistill: Simulated multi-modal distillation for bev 3d object detection," *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38, 2024, pp. 7460–7468.

[30] Gao, Y., Huang, J., Sun, X., Jie, Z., Zhong, Y., and Ma, L., "Matten: Video generation with mamba-attention," *arXiv preprint arXiv:2405.03025*, 2024.

[31] Huang, J., Yan, M., Chen, S., Huang, Y., and Chen, S., "Magicfight: Personalized martial arts combat video generation," *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10833–10842.

[32] Zhang, X., Li, R., Yu, J., Xu, Y., Li, W., and Zhang, J., "Editguard: Versatile image watermarking for tamper localization and copyright protection," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 11964–11974.

[33] Zhang, X., Tang, Z., Xu, Z., Li, R., Xu, Y., Chen, B., Gao, F., and Zhang, J., "Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking," *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 3008–3018.

[34] Huang, Y., Huang, J., Liu, J., Yan, M., Dong, Y., Lv, J., Chen, C., and Chen, S., "Wavedm: Wavelet-based diffusion models for image restoration," *IEEE Transactions on Multimedia*, Vol. 26, 2024, pp. 7058–7073.

[35] Wang, Q., Hu, H., and Zhou, Y., "Memorymamba: Memory-augmented state space model for defect recognition," *arXiv preprint arXiv:2405.03673*, 2024.

[36] Chen, Z., Shen, Y., Song, Y., and Wan, X., "Cross-modal Memory Networks for Radiology Report Generation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 5904–5914. https://doi.org/10.18653/v1/2021.acl-long.459.

[37] Luu, K., Khashabi, D., Gururangan, S., Mandyam, K., and Smith, N. A., "Time Waits for No One! Analysis and Challenges of Temporal Misalignment," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2022, pp. 5944–5958. https://doi.org/10.18653/v1/2022.naacl-main.435.

[38] Aghajanyan, A., Gupta, A., Shrivastava, A., Chen, X., Zettlemoyer, L., and Gupta, S., "Muppet: Massive Multi-task Representations with Pre-Finetuning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2021, pp. 5799–5811. https://doi.org/10.18653/v1/2021.emnlp-main.468.

[39] Wang, Y., Wang, C., Li, R., and Lin, H., "On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2022, pp. 3416–3425. https://doi.org/10.18653/v1/2022.naacl-main.249.

[40] Pan, L., Chen, W., Xiong, W., Kan, M.-Y., and Wang, W. Y., "Unsupervised Multi-hop Question Answering by Question Generation," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2021, pp. 5866–5880. https://doi.org/10.18653/v1/2021.naacl-main.469.

[41] Kryscinski, W., Rajani, N., Agarwal, D., Xiong, C., and Radev, D., "BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization," *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, 2022, pp. 6536–6558. https://doi.org/10.18653/v1/2022.findings-emnlp.488.