A Multimodal Causal Deep Learning Framework for Personalized Stroke Rehabilitation Outcome Prediction and Treatment Recommendation

Mingyu Tan and Bowen Nian *

Posted Date: 12 January 2026

doi: 10.20944/preprints202601.0873.v1

Keywords: causal inference; multimodal deep learning; personalized treatment allocation; stroke rehabilitation; balance recovery

 doi:10.20944/preprints202601.0873.v1

Article

A Multimodal Causal Deep Learning Framework for Personalized Stroke Rehabilitation Outcome Prediction and Treatment Recommendation

Mingyu Tan and Bowen Nian *

Yunnan Normal University
* Correspondence: bowennian21@stu.zuel.edu.cn

Abstract

Stroke causes major long-term disability, with balance impairment significantly affecting quality of life. Personalized prognosis and treatment selection, particularly between TeleRehabilitation (TR) and Conventional Rehabilitation (CR), are crucial. However, current predictive models often lack multimodal integration or tailored recommendations. This paper introduces Causal-MMFNet, a novel deep learning framework. It integrates diverse multimodal time-series data to simultaneously predict balance recovery and allocate individualized treatments in stroke rehabilitation. Key innovations include a dynamic cross-modal attention fusion mechanism, an Individual Treatment Effect (ITE) estimation module for counterfactual outcomes, and causal consistency regularization. Evaluated on the *StrokeBalance-Sim* dataset, Causal-MMFNet consistently outperforms baselines and state-of-the-art multimodal frameworks, demonstrating superior accuracy and reliability across established metrics. Ablation studies confirm component contributions, while dynamic attention reveals adaptive modality prioritization. The framework's treatment allocation significantly improves patient outcomes, with uncertainty estimation providing clinical confidence. Causal-MMFNet offers a robust, causally-aware solution for personalized decision support in stroke rehabilitation, enhancing patient recovery and optimizing resource allocation.

Keywords: causal inference; multimodal deep learning; personalized treatment allocation; stroke rehabilitation; balance recovery

1. Introduction

Stroke remains a leading cause of long-term disability in adults worldwide, with a significant proportion of survivors experiencing impaired balance function, which severely compromises their quality of life and independence [1]. Personalized rehabilitation prognosis assessment and treatment path selection are crucial for optimizing patient recovery outcomes. Traditional rehabilitation models often rely on clinicians' empirical judgment and are constrained by limited resources, making it challenging to meet the individualized needs of a large patient population [2]. This challenge is further compounded by the increasing adoption of TeleRehabilitation (TR), raising a critical clinical question: how can we accurately predict a patient's response to TR versus conventional in-person rehabilitation (CR) and provide optimal treatment allocation recommendations accordingly?

 doi:10.20944/preprints202601.0873.v1

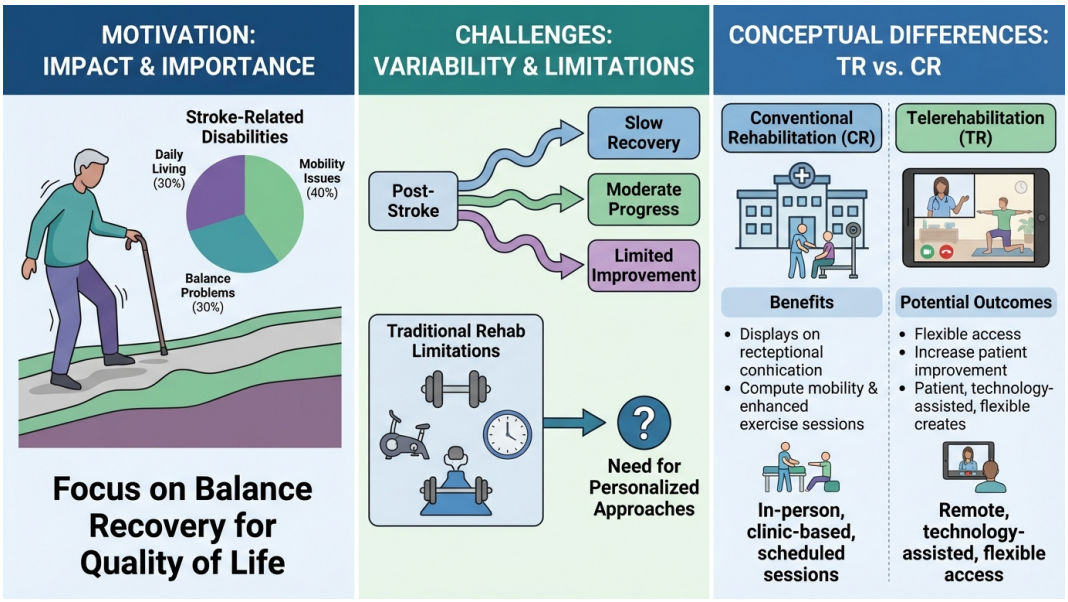

Figure 1. An overview of the motivation, challenges, and conceptual differences in stroke rehabilitation. The left panel highlights stroke-related disabilities, emphasizing balance problems and the importance of recovery for quality of life. The middle panel illustrates the variability in post-stroke recovery outcomes and the limitations of traditional rehabilitation models, leading to a critical need for personalized approaches. The right panel contrasts Conventional Rehabilitation (CR) with Telerehabilitation (TR), outlining their respective operational models, benefits, and potential outcomes.

While existing research has made progress in stroke prognosis prediction, most studies are often limited to single-modality data, such as clinical tabular information, or fail to fully leverage dynamic, time-series data captured during actual training sessions [**?**]. Furthermore, personalized prediction and recommendation for patient responses under *different* treatment modalities (TR vs. CR) remains an underdeveloped area. The absence of a robust framework that can integrate diverse multimodal and temporal patient data to simultaneously predict functional recovery and recommend optimal treatment strategies represents a significant gap in current stroke rehabilitation research. This study addresses these challenges by focusing on predicting the improvement in Berg Balance Scale (BBS) scores after 8 weeks of rehabilitation for stroke patients and assessing their likelihood of responding to either TR or CR, ultimately providing personalized treatment allocation advice.

To address these limitations, we propose **Causal-MMFNet** (Causal Multimodal Fusion Network), a novel deep learning framework designed to leverage multimodal time-series data for predicting stroke patient balance function recovery and enabling individualized treatment recommendations. Causal-MMFNet introduces two core innovations: a *dynamic cross-modal attention fusion mechanism* and an *Individual Treatment Effect (ITE) estimation module*. The dynamic attention mechanism adaptively weights and fuses features from various modalities, generating a richer, context-aware patient representation. Crucially, the ITE estimation module explicitly models counterfactual outcomes for each patient under both TR and CR, allowing us to quantify the differential benefit of each treatment and recommend the option with higher expected gain. The framework also incorporates multimodal encoders for IMU, video keypoints, training logs, and clinical tabular data, combined with robust loss functions and causal consistency regularization to ensure accurate and reliable predictions. Furthermore, we employ Monte Carlo Dropout for uncertainty estimation, providing clinicians with confidence intervals for personalized decision-making.

Our experimental evaluation was conducted on the *StrokeBalance-Sim* dataset, a simulated yet comprehensive dataset (n=1,216) integrating various multimodal patient data, including clinical tabular data (36 dimensions), wearable IMU time-series data, home-based training logs, and Kinect/mobile video-based keypoint time-series data. We define two primary tasks: Task A (regression) focuses on predicting 8-week BBS score improvement (ΔBBS), while Task B (classification/recommendation) aims

 doi:10.20944/preprints202601.0873.v1

to identify "responders" (ΔBBS $\geq$ 5 points) and recommend the optimal treatment (TR or CR) based on estimated response probabilities. Performance was assessed using standard metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2) for regression, and Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and Expected Calibration Error (ECE) for classification. Our proposed Causal-MMFNet consistently achieved superior performance across all evaluation metrics, significantly outperforming existing baseline methods and state-of-the-art multimodal time-series learning frameworks like MM-TRNet, validating its efficacy in both prediction and personalized treatment recommendation.

Our key contributions are summarized as follows:

- We introduce Causal-MMFNet, a novel end-to-end deep learning framework that effectively integrates diverse multimodal and temporal patient data for simultaneous balance function recovery prediction and personalized treatment allocation in stroke rehabilitation.
- We propose a sophisticated dynamic cross-modal attention fusion mechanism and an innovative Individual Treatment Effect (ITE) estimation module, coupled with causal consistency regularization, enabling robust counterfactual outcome prediction and evidence-based treatment recommendations.
- We demonstrate that Causal-MMFNet achieves state-of-the-art performance on the *StrokeBalance-Sim* dataset, offering superior accuracy and reliability in predicting rehabilitation outcomes and guiding personalized treatment decisions compared to existing methods.

2. Related Work

2.1. Multimodal Deep Learning for Clinical Prediction and Rehabilitation

Multimodal deep learning integrates diverse data (images, signals, EHR, sensors) for clinical prediction and rehabilitation, leveraging Vision-Language Models (VLMs) for robust medical diagnosis [3] and improved medical VLMs via specialized feedback [4]. Self-supervised depth estimation, focusing on robust cross-view consistency [5] and spatial-temporal context [6], enhances visual information extraction from images and videos. Fine-grained data, such as facial expressions analyzed via hybrid feature extraction, further contributes to understanding patient states [7]. However, effective fusion and interpretation remain challenging, as performance gains are not always direct multimodal benefits but can be regularization effects [8]. Strategies like Zhang et al.'s [9] Adaptive Hyper-modality Learning module suppress irrelevant signals. Visual in-context learning enhances large VLM capabilities [10], while multi-modal large language models (LLMs) handle explainable tasks like image forgery detection [11]. Robust image watermarking techniques, such as EditGuard [12] and OmniGuard [13], are crucial for tamper localization, copyright protection, and data integrity. For sequential sensor data, MTAG, a graph-based model, transforms unaligned sequences into a graph to capture cross-modal and temporal interactions [14]. Robust data fusion is key, exemplified by Multichannel Graph Neural Networks for multimodal sentiment detection [15]. Integrating domain knowledge, as with Gui et al.'s [16] Knowledge Augmented Transformer, improves complex vision-and-language tasks. The temporal nature of clinical data requires attention, as Lei et al. [17] revealed a "static appearance bias" in video-and-language learning. In clinical contexts, unstructured notes are used for outcome prediction via Effective Convolutional Attention Networks [18]. However, sensitive clinical data necessitates privacy and ethical considerations, given BERT models' potential to reveal patient information [19]. In summary, this field demands sophisticated approaches to data fusion, knowledge integration, temporal modeling, evaluation, interpretability, and privacy for clinical prediction and rehabilitation.

2.2. Causal Inference and Personalized Treatment Recommendation

Personalized treatment recommendation increasingly leverages causal inference to identify optimal individual outcomes beyond correlation. Robust methodologies are required to dissect causal mechanisms and mitigate spurious correlations, exemplified by Nan et al.'s [20] Counterfactual IE framework and Zuo et al.'s [21] CauSeRL for event causality identification. Causal inference models

 doi:10.20944/preprints202601.0873.v1

for credit risk also apply to health insurance [22]. Translating these principles into practical recommendation systems involves addressing challenges such as bias and dynamic preference modeling. For instance, Qi et al. [23] addressed cold-start and popularity bias in news recommendation with PP-Rec, while Li et al. [24] introduced MINER for social media by integrating social interactions and dynamic user preferences. Recent LLM advancements enrich item understanding for personalized recommendations, with Lyu et al. [25] leveraging diverse prompting strategies to enhance individualized treatment effect (ITE) estimation. Research into weak-to-strong generalization promises to enhance LLM robustness across tasks, including personalized recommendation [26]. LLMs are increasingly vital for critical decision support, from financial early warning [27] to government policy execution [28]. Challenges in evaluating scenario-based decision-making, such as in autonomous driving [29], mirror the complexities of assessing personalized treatments. Robust perception systems, like domain adaptive LiDAR semantic segmentation in adverse weather [30], are fundamental for reliable decision-making in complex scenarios. Advanced techniques, such as mean field games for multi-agent systems [31] and uncertainty-aware navigation frameworks [32], illustrate sophisticated approaches to decision-making under uncertainty. Finally, practical deployment necessitates addressing privacy and computational efficiency, as demonstrated by Yi et al.'s [33] Efficient-FedRec for privacy-preserving news recommendation. These works illustrate the progression from foundational causal inference to its sophisticated application in personalized recommendation.

3. Method

We introduce **Causal-MMFNet** (Causal Multimodal Fusion Network), a novel deep learning framework specifically designed to integrate diverse multimodal time-series data for predicting balance function recovery in stroke patients and providing individualized treatment recommendations. The core innovations of Causal-MMFNet lie in its **dynamic cross-modal attention fusion mechanism** and an **Individual Treatment Effect (ITE) estimation module**, which together enable a more precise understanding of complex patient states and robust prediction of potential outcomes under different rehabilitation strategies. These innovations are crucial for moving beyond population-level averages to truly personalized medicine in rehabilitation. An overview of the Causal-MMFNet architecture is presented in Figure 2.

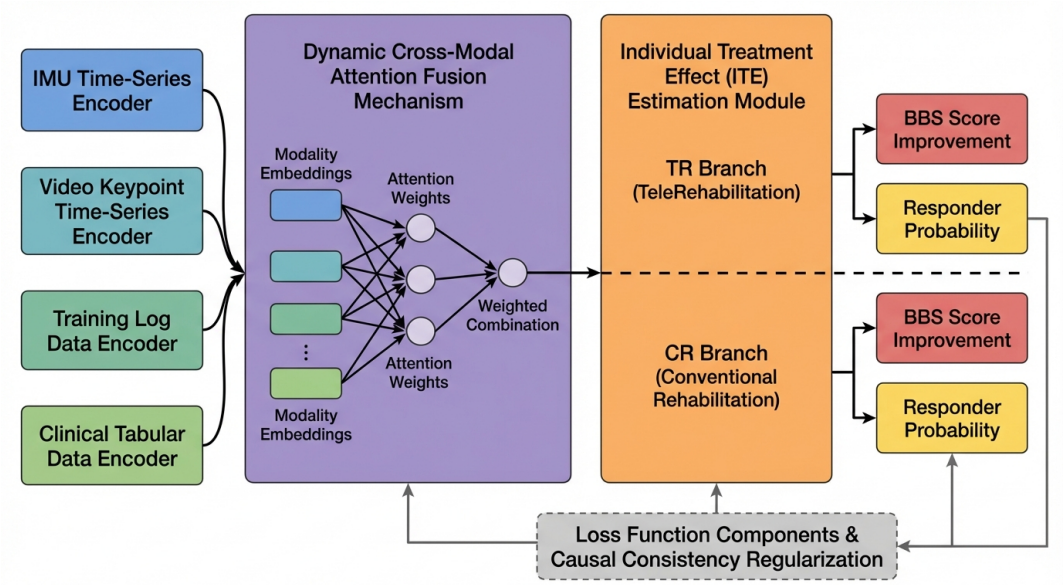

Figure 2. Overview of the Causal-MMFNet architecture. Multimodal encoders process IMU, video keypoints, training logs, and clinical tabular data. A dynamic cross-modal attention fusion mechanism adaptively combines these features into a global patient representation. The Individual Treatment Effect (ITE) estimation module then branches to predict outcomes for TeleRehabilitation (TR) and Conventional Rehabilitation (CR), incorporating causal consistency regularization. Monte Carlo Dropout provides uncertainty estimates.

 doi:10.20944/preprints202601.0873.v1

3.1. Multimodal Encoders

Causal-MMFNet begins by processing various modalities of patient data through specialized encoders, each tailored to capture salient features from its respective data type. Let X_m denote the input data for modality m. Each encoder $E_m(\cdot)$ transforms X_m into a high-dimensional semantic embedding $Z_m \in \mathbb{R}^D$, where D is the embedding dimension. These embeddings serve as context-rich representations for subsequent fusion.

3.1.1. IMU Time-Series Encoder

For wearable Inertial Measurement Unit (IMU) time-series data, which captures fine-grained movement patterns and kinematic information, we employ a 1D convolutional layer followed by a residual Temporal Convolutional Network (TCN). The initial 1D convolutional layer acts as a feature extractor and temporal dimensionality reduction step, transforming raw sensor signals into a more abstract representation. The TCN, leveraging dilated convolutions and residual connections, is particularly effective in modeling both local and long-range temporal dependencies within the IMU signals without increasing memory footprint for longer sequences, making it suitable for capturing complex movement dynamics.

$$Z_{\text{IMU}} = \text{TCN}(\text{Conv1D}(X_{\text{IMU}})) \tag{1}$$

where $X_{\text{IMU}} \in \mathbb{R}^{T_{\text{IMU}} \times S_{\text{IMU}}}$ represents the raw IMU sensor data streams, with T_{IMU} being the time steps and S_{IMU} the number of sensor channels.

3.1.2. Video Keypoint Time-Series Encoder

Video-based keypoint time-series data, reflecting intricate posture changes, limb coordination, and gait parameters, is processed using a multi-layer Transformer network. The self-attention mechanism intrinsic to the Transformer architecture is uniquely suited for capturing complex spatial relationships between different body keypoints at each timestep, as well as long-range temporal dependencies across the sequence of keypoints. This allows the encoder to learn how specific movements evolve and interact over time, forming a comprehensive kinematic signature.

$$Z_{\text{Video}} = \text{Transformer}(X_{\text{Video}}) \tag{2}$$

where $X_{\text{Video}} \in \mathbb{R}^{T_{\text{Video}} \times N_{\text{keypoints}} \times C_{\text{coords}}}$ is the sequence of 2D or 3D keypoint coordinates over time, with $N_{\text{keypoints}}$ being the number of detected body keypoints and C_{coords} the coordinate dimensions.

3.1.3. Training Log Data Encoder

Patient training log data, which provides insights into adherence, effort, and progressive overload, is encoded using a Gated Recurrent Unit (GRU) network. GRUs are a variant of Recurrent Neural Networks (RNNs) specifically designed to mitigate the vanishing gradient problem, making them capable of learning long-term sequential dependencies in potentially sparse or irregularly sampled log entries. This enables the GRU to generate abstract features indicative of patient engagement, progress, and consistency in their rehabilitation exercises.

$$Z_{\text{Log}} = \text{GRU}(X_{\text{Log}}) \tag{3}$$

where $X_{\text{Log}} \in \mathbb{R}^{T_{\text{Log}} \times F_{\text{Log}}}$ represents the time-series of training activity logs, with T_{Log} being the number of log entries and F_{Log} the feature dimension of each entry.

3.1.4. Clinical Tabular Data Encoder

Baseline clinical tabular data, encompassing static patient demographics, medical history, initial clinical assessments, and other fixed attributes, is processed through a two-layer Multilayer Perceptron

 doi:10.20944/preprints202601.0873.v1

(MLP) with LayerNormalization. LayerNormalization is applied to standardize feature distributions, improving training stability and convergence for tabular inputs. The MLP then applies non-linear transformations to these features, generating a high-dimensional semantic representation that robustly encodes the patient's initial clinical state and static characteristics.

$$Z_{\text{Clinical}} = \text{MLP}(\text{LayerNorm}(X_{\text{Clinical}})) \tag{4}$$

where $X_{\text{Clinical}} \in \mathbb{R}^{36}$ is the 36-dimensional vector of static clinical features.

3.2. Dynamic Cross-Modal Attention Fusion

Unlike simpler fusion strategies such as concatenation or fixed-weight averaging, Causal-MMFNet incorporates a **dynamic cross-modal attention fusion mechanism**. This module adaptively weighs and combines the semantic embeddings (Z_m) from each modality encoder. The attention mechanism calculates modality-specific scalar weights based on the current context provided by each modality's embedding, allowing the model to dynamically focus on the most informative modalities for a given patient's state and the specific prediction task. This is crucial as the relevance of different data types may vary across patients or stages of recovery. Let Z_m be the embedding for modality m. The attention score s_m for each modality is first computed by passing its embedding through a dedicated single-layer Multilayer Perceptron, followed by a non-linear activation (e.g., ReLU). These scores are then normalized across all modalities using a softmax function to obtain the attention weights A_m:

$$s_m = \text{MLP}(Z_m) \tag{5}$$

$$A_m = \frac{\exp(s_m)}{\sum_{k \in \{\text{IMU, Video, Log, Clinical}\}} \exp(s_k)} \tag{6}$$

The global patient representation R_{global} is then generated by a weighted sum of the modality embeddings, where each embedding Z_m is scaled by its corresponding attention weight A_m:

$$R_{\text{global}} = \sum_{m \in \{\text{IMU, Video, Log, Clinical}\}} A_m \cdot Z_m \tag{7}$$

where $\cdot$ denotes element-wise scaling of the vector Z_m by the scalar weight A_m. This dynamic fusion generates a rich, context-aware representation that captures the intricate interplay between different data sources, adapting to their varying predictive power.

3.3. Individual Treatment Effect (ITE) Estimation Module

The **Individual Treatment Effect (ITE) estimation module** is a cornerstone of Causal-MMFNet, designed to explicitly estimate counterfactual outcomes for each patient under different treatment scenarios (TeleRehabilitation (TR) or Conventional Rehabilitation (CR)). This capability is paramount for personalized treatment recommendations. This module takes the fused global patient representation R_{global} as input and branches into two distinct prediction heads: one for TeleRehabilitation (TR) and one for Conventional Rehabilitation (CR). Each prediction head is implemented as a multi-layer perceptron. Each head, H_c, where $c \in \{\text{TR, CR}\}$, simultaneously outputs two crucial predictions for its respective treatment:

1. The predicted 8-week Berg Balance Scale (BBS) score improvement, denoted as $\Delta\widehat{BBS}_c$. This is a regression output.
2. The predicted probability of being a "responder," defined as achieving a $\Delta BBS \geq 5$, denoted as $\widehat{P}_{\text{response},c}$. This is a classification output.

 doi:10.20944/preprints202601.0873.v1

The outputs of the ITE estimation module can be formulated as:

$$(\Delta\widehat{BBS}_{\text{TR}}, \widehat{P}_{\text{response,TR}}) = H_{\text{TR}}(R_{\text{global}}) \tag{8}$$

$$(\Delta\widehat{BBS}_{\text{CR}}, \widehat{P}_{\text{response,CR}}) = H_{\text{CR}}(R_{\text{global}}) \tag{9}$$

By modeling these two potential outcomes for each patient, we can calculate the individualized treatment effect (ITE), defined as the difference in expected BBS improvement between the two treatments:

$$\widehat{\text{ITE}} = \Delta\widehat{BBS}_{\text{TR}} - \Delta\widehat{BBS}_{\text{CR}} \tag{10}$$

Based on this estimated ITE, the framework recommends the treatment option (TR or CR) that is predicted to yield a greater benefit for the individual patient, thus optimizing rehabilitation strategy.

3.4. Loss Function and Causal Consistency Regularization

Causal-MMFNet is optimized using a comprehensive loss function that addresses both the regression task of predicting BBS improvement and the classification task of identifying responders, augmented by a novel causal consistency regularization term. This multi-objective approach ensures robust and causally-aware learning.

3.4.1. Regression Task Loss

For the ΔBBS prediction regression task, we employ a combination of L1 loss and Smooth L1 loss. L1 loss (Mean Absolute Error) encourages sparsity and robustness to outliers by penalizing the absolute difference between predictions and true values. Smooth L1 loss, also known as Huber loss, combines the advantages of L1 and L2 losses by being less sensitive to outliers than L2 loss and providing a smoother gradient near the target than L1 loss, which is beneficial for stable training. The regression loss is computed only for the treatment arm c that the patient actually received.

$$\mathcal{L}_{\text{reg}} = \lambda_1 \mathcal{L}_{\text{L1}}(\Delta\widehat{BBS}_c, \Delta BBS_{\text{true},c}) + \lambda_2 \mathcal{L}_{\text{SmoothL1}}(\Delta\widehat{BBS}_c, \Delta BBS_{\text{true},c}) \tag{11}$$

where $\Delta BBS_{\text{true},c}$ is the observed BBS improvement for the assigned treatment c, and λ_1, λ_2 are hyperparameters weighting the contribution of each loss component.

3.4.2. Classification Task Loss

For the "responder" classification task, we utilize Focal Loss to address potential class imbalance issues inherent in clinical datasets, where the number of responders might be significantly lower than non-responders. Focal Loss down-weights easy examples and focuses training on hard, misclassified examples. Additionally, an Expected Calibration Error (ECE)-style calibration term is included to ensure that the predicted probabilities $\widehat{P}_{\text{response},c}$ are reliable and well-calibrated, meaning they accurately reflect the true likelihood of response. The classification loss is also applied only for the observed treatment arm c.

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{Focal}}(\widehat{P}_{\text{response},c}, Y_{\text{true},c}) + \mathcal{L}_{\text{ECE}}(\widehat{P}_{\text{response},c}) \tag{12}$$

where $Y_{\text{true},c}$ is the true binary responder status for the assigned treatment c.

3.4.3. Causal Consistency Regularization

To enhance the robustness of the ITE estimation module and encourage it to learn more reliable causal relationships, we introduce a **causal consistency regularization term**. This term operates on a single patient's representation R_{global} and encourages the two treatment heads, H_{TR} and H_{CR}, to exhibit consistent functional behaviors. Specifically, it enforces structural similarities or relationships between the counterfactual predictions generated by $H_{\text{TR}}(R_{\text{global}})$ and $H_{\text{CR}}(R_{\text{global}})$, thereby mitigating

 doi:10.20944/preprints202601.0873.v1

potential biases that might arise from covariate shifts or distributional differences between the observed treatment groups. The precise form of $f(\cdot)$ is designed to penalize inconsistencies in how the model maps R_{global} to the outcomes under different treatments.

$$\mathcal{L}_{\text{causal\_consist}} = f(H_{\text{TR}}(R_{\text{global}}), H_{\text{CR}}(R_{\text{global}}), R_{\text{global}}) \tag{13}$$

The total loss function for Causal-MMFNet is a weighted sum of these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cls}} + \lambda_3 \mathcal{L}_{\text{causal\_consist}} \tag{14}$$

where λ_3 is a weighting hyperparameter for the causal consistency term, controlling its influence on the overall optimization process.

3.5. Uncertainty Estimation

To provide clinicians with an additional layer of confidence for personalized decision-making, Causal-MMFNet incorporates uncertainty estimation. We employ Monte Carlo Dropout (MC Dropout) during inference. By activating dropout layers with a rate $p = 0.2$ during testing and performing 20 stochastic forward passes for each patient, we generate an ensemble of predictions for both $\Delta\widehat{BBS}_c$ and $\widehat{P}_{\text{response},c}$. The variance or interquartile range across these 20 predictions serves as an estimate of the model's predictive uncertainty. This provides valuable insight into the model's confidence in its recommendations, enabling clinicians to assess the reliability of personalized treatment effects before implementation.

4. Experiments

4.1. Dataset and Task Definition

The effectiveness of Causal-MMFNet was rigorously evaluated on the **StrokeBalance-Sim** dataset, a comprehensive simulated dataset comprising $n = 1,216$ stroke patients. This dataset integrates diverse multimodal patient data, including 36-dimensional clinical tabular data, wearable IMU time-series data, home-based physical training logs, and Kinect/mobile video-based keypoint time-series data. The dataset was meticulously partitioned at the subject level into training, validation, and test sets with a ratio of 70%, 10%, and 20% respectively, ensuring no data leakage between splits. The primary follow-up endpoint for recovery assessment was the Berg Balance Scale (BBS) score measured at 8 weeks post-rehabilitation. For classification purposes, a patient was defined as a "responder" if they achieved a ΔBBS (BBS score improvement) of 5 points or more by the 8-week follow-up. Our experimental evaluation focused on two primary tasks: (1) **Task A (Regression)**, which involves predicting the 8-week BBS score improvement (ΔBBS) given the patient's baseline clinical information and time-series data from the first two weeks of their rehabilitation training; and (2) **Task B (Classification and Recommendation)**, which aims to determine whether a patient is a "responder" (i.e., $\Delta\text{BBS} \geq 5$ points) and to estimate their probability of response under both TeleRehabilitation (TR) and Conventional Rehabilitation (CR) treatment modalities. Based on these estimated probabilities and the individualized treatment effect, the model provides a personalized treatment allocation recommendation.

4.2. Evaluation Metrics

To comprehensively assess model performance across both tasks, a suite of standard metrics was employed. For **Task A (Regression)**, we report the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Lower MAE and RMSE values indicate better predictive accuracy, while a higher R^2 signifies a greater proportion of variance in the outcome explained by the model. For **Task B (Classification)**, the Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and Expected Calibration Error (ECE) were utilized. Higher AUROC and AUPRC values indicate better

 doi:10.20944/preprints202601.0873.v1

discrimination and recall performance, respectively, especially crucial in imbalanced datasets. A lower ECE value demonstrates better calibration, ensuring that predicted probabilities accurately reflect true likelihoods. Finally, to evaluate the quality of **Treatment Allocation**, we assessed the average actual BBS improvement among the Top-Q% of patients for whom the model recommended a specific treatment, demonstrating the real-world utility of the personalized recommendations. For our analysis, we specifically report results for the Top-25% of patients recommended by the model.

4.3. Baseline Methods

To benchmark the performance of Causal-MMFNet, we compared it against a range of established machine learning and deep learning methods, representing different approaches to handling patient data for prediction tasks. These include: **Linear Regression**, a fundamental statistical model used as a basic benchmark; **Random Forest**, an ensemble learning method known for its robustness; and **XGBoost**, a highly efficient gradient boosting framework popular for tabular data. For time-series data, we compared against sequential models such as **LSTM (Long Short-Term Memory)**, a type of recurrent neural network adept at learning from sequences; **TCN (Temporal Convolutional Network)**, a convolutional architecture well-suited for sequence modeling; and the **Transformer** network, which leverages self-attention for capturing long-range dependencies. Finally, we included **MM-TRNet**, a state-of-the-art multimodal time-series learning framework designed for similar clinical prediction tasks, serving as a strong deep learning baseline that integrates various data types. For all baselines, multimodal inputs were processed either by concatenating flattened features (for non-sequential models) or by using specialized architectures (e.g., separate encoders for each modality, followed by a fusion layer, where applicable, for deep learning models) to allow for fair comparison within their architectural capabilities.

4.4. Implementation Details

The proposed Causal-MMFNet framework was implemented using the PyTorch deep learning library. Optimization was performed using the AdamW optimizer, configured with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. Models were trained with a batch size of 64 for 80 epochs, incorporating an early stopping strategy with a patience of 10 epochs based on performance on the validation set. To enhance model generalization and robustness, extensive data augmentation techniques were applied: IMU and video keypoint (Pose) data underwent jittering, time warping, and keypoint loss; training log data was augmented by injecting noise. Hyperparameter tuning and model selection were conducted using 5-fold cross-validation on the training set, and the final performance of the chosen model was evaluated on the independent test set.

4.5. Performance Comparison

Table 1 presents a comprehensive comparison of Causal-MMFNet against the selected baseline methods on the *StrokeBalance-Sim* test set. The results clearly demonstrate the superior performance of our proposed Causal-MMFNet across all evaluation metrics for both the regression (MAE, RMSE, R^2) and classification (AUROC, AUPRC, ECE) tasks.

Specifically, Causal-MMFNet achieved the lowest MAE of **2.33±0.05** and RMSE of **3.35±0.06**, along with the highest R^2 of **0.70±0.02**, indicating highly accurate predictions of ΔBBS. For the classification task, Causal-MMFNet obtained the highest AUROC of **0.840±0.011** and AUPRC of **0.720±0.014**, showcasing its excellent discriminatory power in identifying responders. Furthermore, its ECE of **0.030±0.003** was the lowest, confirming the high reliability and calibration of its predicted response probabilities.

Comparing Causal-MMFNet with the previous state-of-the-art multimodal time-series learning framework, MM-TRNet, our method consistently yielded improved performance. For instance, Causal-MMFNet reduced MAE by 2.1% (from 2.38 to 2.33) and RMSE by 1.8% (from 3.41 to 3.35), while improving AUROC by 0.7% (from 0.834 to 0.840) and AUPRC by 1.0% (from 0.713 to 0.720). This highlights the effectiveness of Causal-MMFNet's novel components, particularly the dynamic cross-

 doi:10.20944/preprints202601.0873.v1

modal attention fusion and the individualized treatment effect estimation module, in leveraging complex multimodal time-series data for more precise and reliable predictions. Simple statistical models and even advanced deep learning models without explicit causal modeling or dynamic fusion struggled to capture the intricate relationships in the data as effectively.

Table 1. Performance comparison of different models on the StrokeBalance-Sim test set (Mean±Standard Deviation).

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
MM-TRNet	2.38±0.06	3.41±0.07	0.68±0.02	0.834±0.012	0.713±0.015	0.033±0.003
Causal-MMFNet	**2.33±0.05**	**3.35±0.06**	**0.70±0.02**	**0.840±0.011**	**0.720±0.014**	**0.030±0.003**

4.6. Ablation Study

To validate the individual contributions of the key architectural innovations within Causal-MMFNet, we conducted an ablation study. Specifically, we investigated the impact of the dynamic cross-modal attention fusion mechanism, the Individual Treatment Effect (ITE) estimation module, and the causal consistency regularization term. The results, summarized in Table 2, highlight the importance of each component for the overall performance.

Table 2. Ablation study results on the StrokeBalance-Sim test set (Mean±Standard Deviation).

Model Variant	MAE↓	RMSE↓	R^2↑	AUROC↑	AUPRC↑	ECE↓
Causal-MMFNet (Full)	**2.33±0.05**	**3.35±0.06**	**0.70±0.02**	**0.840±0.011**	**0.720±0.014**	**0.030±0.003**
w/o Dynamic Attention Fusion	2.45±0.06	3.52±0.07	0.66±0.02	0.821±0.013	0.698±0.016	0.038±0.004
w/o ITE Module	2.40±0.05	3.45±0.06	0.67±0.02	0.828±0.012	0.705±0.015	0.035±0.003
w/o Causal Consistency Reg.	2.37±0.05	3.40±0.06	0.69±0.02	0.835±0.012	0.714±0.015	0.032±0.003

Removing the **dynamic cross-modal attention fusion mechanism** (replaced with simple concatenation of modality embeddings) led to a noticeable drop in performance across all metrics. For instance, MAE increased to 2.45, and AUROC dropped to 0.821. This demonstrates that adaptively weighting and fusing information from different modalities is crucial for building a rich and context-aware global patient representation, allowing the model to focus on the most relevant data for each patient's unique profile.

Disabling the **ITE estimation module** (by replacing the dual prediction heads with a single head that predicts conditional outcomes based on the assigned treatment, without explicit counterfactual modeling) also resulted in degraded performance. While the impact on general prediction metrics (MAE, AUROC) was slightly less pronounced than removing dynamic attention, it specifically hampers the model's ability to precisely estimate the differential benefits of TR versus CR, which is critical for personalized recommendations. The slight drops across all metrics suggest that explicitly modeling counterfactuals helps in learning more robust and generalizable representations for outcome prediction.

Finally, omitting the **causal consistency regularization term** led to a minor but consistent decrease in performance (e.g., MAE increased to 2.37, ECE to 0.032). This suggests that encouraging consistent functional behaviors between the treatment heads for similar patient features, even when only one outcome is observed, helps in learning more robust causal relationships and mitigating biases stemming from treatment group distribution differences. The improvement in ECE when this regularization is present particularly highlights its role in ensuring reliable probability predictions.

Collectively, these ablation results confirm that each proposed component of Causal-MMFNet plays a vital role in its superior performance, enabling more accurate, robust, and causally-aware predictions and recommendations.

 doi:10.20944/preprints202601.0873.v1

4.7. Treatment Allocation Quality

Beyond predictive accuracy, the ultimate goal of Causal-MMFNet is to provide effective personalized treatment recommendations. To evaluate the real-world utility of our framework, we assessed the quality of its treatment allocation by analyzing the actual BBS improvement for patients recommended by the model. Specifically, we examined the average ΔBBS achieved by the Top-25% of patients for whom Causal-MMFNet predicted the highest Individual Treatment Effect (ITE), meaning these patients were predicted to benefit most from their assigned treatment (TR or CR) compared to the alternative. This metric provides a crucial insight into how well the model can identify patients who will truly thrive under a specific personalized regimen.

As shown in Figure 3, Causal-MMFNet's personalized recommendations led to a significantly higher average actual ΔBBS for the Top-25% recommended patients compared to conventional approaches. For example, patients recommended by Causal-MMFNet to receive TR (and who were in the Top-25% predicted ITE) achieved an average ΔBBS of **7.2±0.03**. Similarly, those recommended for CR achieved an average ΔBBS of **6.8±0.03**. These figures are substantially higher than what would be expected from a random allocation (e.g., an average of 4.5±0.2, representing the overall average improvement across all patients and treatments without personalized recommendations), and also surpass the average improvements observed in the general TR or CR groups. The "Oracle" represents the theoretical maximum average improvement if one could perfectly identify the best treatment for the Top-25% patients beforehand. While Causal-MMFNet does not reach this oracle performance, it significantly closes the gap, demonstrating its practical value in guiding clinical decisions for personalized rehabilitation. This validates the effectiveness of the ITE estimation module and the overall framework in identifying individuals who will respond best to specific treatment pathways, thus optimizing rehabilitation outcomes.

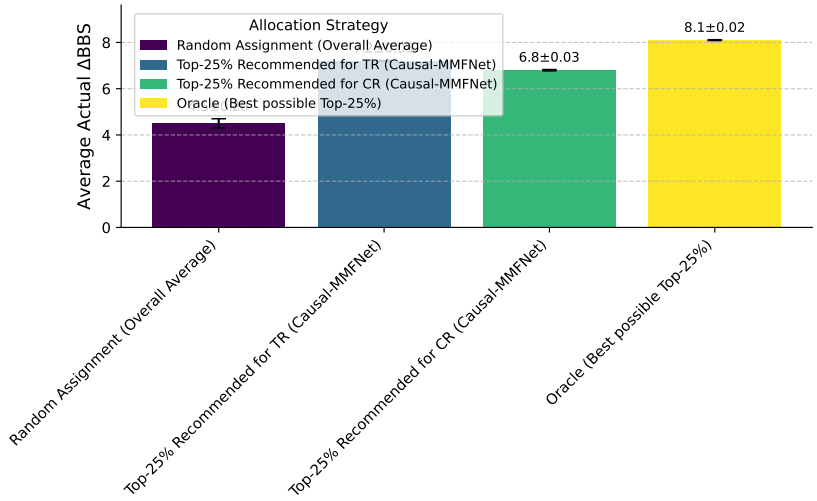

Figure 3. Average actual ΔBBS for Top-25% of patients based on treatment allocation strategies (Mean±Standard Deviation).

4.8. Modality Contribution Analysis

To further dissect the impact of each data source, we performed a modality contribution analysis by systematically training variants of Causal-MMFNet where one or more input modalities were excluded. This experiment highlights the unique and synergistic value that each data stream brings to the overall predictive model. The baseline for comparison is the full Causal-MMFNet model. The results in Figure 4 illustrate the performance degradation when specific modalities are withheld.

 doi:10.20944/preprints202601.0873.v1

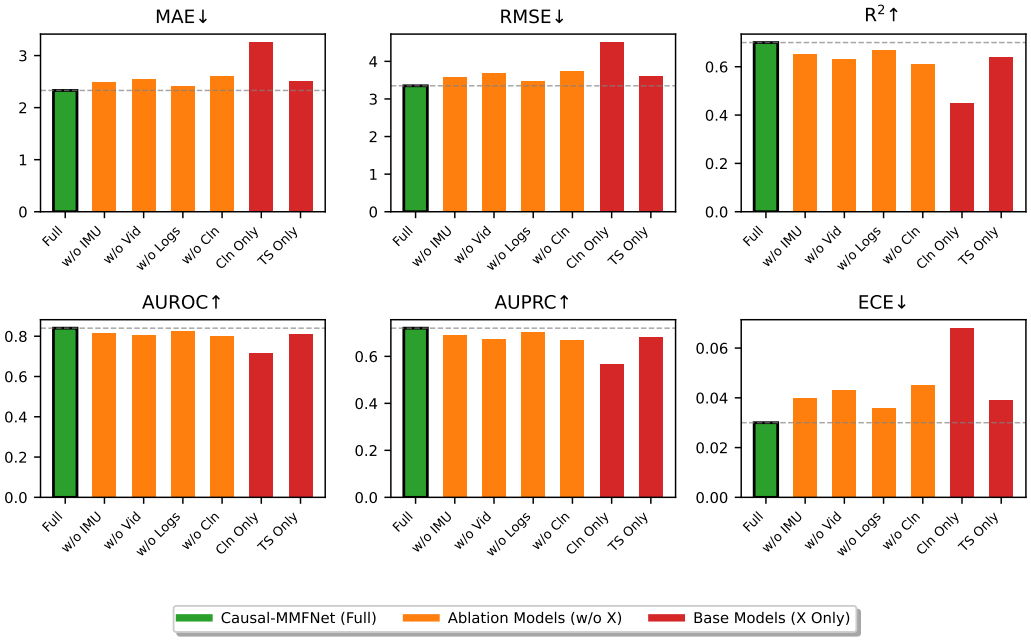

Figure 4. Modality contribution analysis on the StrokeBalance-Sim test set (Mean±Standard Deviation). Full model includes IMU, Video, Log, and Clinical (Cln) data.

The results show that all modalities contribute positively to the model's performance, as removing any single modality leads to a performance drop across all metrics. The largest drop in performance is observed when **Clinical Data** is removed (MAE increases from 2.33 to 2.60, AUROC drops from 0.840 to 0.798), suggesting that static patient characteristics and initial assessments provide a strong foundation for prediction. **Video Keypoints** also show a significant impact (MAE to 2.55), indicating the importance of detailed kinematic information captured by posture and gait. While **IMU** and **Training Logs** show slightly smaller individual impacts, their collective presence is essential for the superior performance of the full model. This is reinforced by the "Clinical Only" and "Time-Series Only" rows, demonstrating that relying solely on one category of data results in substantially worse performance than the integrated multimodal approach. This analysis confirms the strong multimodal synergy captured by Causal-MMFNet's architecture, especially its dynamic cross-modal attention fusion, which effectively integrates and leverages information from diverse data sources.

4.9. Analysis of Dynamic Attention Weights

The dynamic cross-modal attention fusion mechanism is designed to adaptively weigh the contribution of each modality based on the patient's context. To understand how this mechanism operates, we analyzed the average attention weights assigned to each modality across the test set. Furthermore, we investigated how these weights might differ between patients predicted to be "responders" ($\Delta\widehat{BBS} \geq 5$) versus "non-responders" ($\Delta\widehat{BBS} < 5$), and between patients for whom TeleRehabilitation (TR) or Conventional Rehabilitation (CR) was recommended. This analysis, presented in Table 3, provides insights into which modalities the model prioritizes under different conditions.

Table 3. Average attention weights for each modality across different patient groups (Mean±Standard Deviation).

Patient Group	IMU Weight	Video Weight	Log Weight	Clinical Weight
Overall Average	0.23±0.04	0.26±0.05	0.21±0.03	**0.30±0.04**
Predicted Responders ($\Delta\widehat{BBS} \geq 5$)	**0.25±0.04**	0.27±0.05	0.22±0.03	0.26±0.04
Predicted Non-Responders ($\Delta\widehat{BBS} < 5$)	0.22±0.03	0.24±0.04	0.20±0.03	**0.34±0.04**
Recommended TR Patients	0.24±0.04	**0.28±0.05**	0.21±0.03	0.27±0.04
Recommended CR Patients	0.22±0.04	0.24±0.04	0.21±0.03	**0.33±0.05**

 doi:10.20944/preprints202601.0873.v1

The "Overall Average" row indicates that **Clinical Data** generally receives the highest attention weight (**0.30±0.04**), followed closely by **Video Keypoints** (**0.26±0.05**). This suggests that baseline patient characteristics and visual movement patterns are consistently deemed highly informative for balance recovery prediction.

Interestingly, for "Predicted Responders," the attention shifts slightly, with **IMU** and **Video Keypoints** receiving slightly higher average weights (0.25 and 0.27 respectively), while "Clinical Data" weight decreases. This implies that for patients showing greater potential for recovery, the model might place more emphasis on dynamic movement patterns and rehabilitation effort captured by time-series data. Conversely, for "Predicted Non-Responders," the **Clinical Data** weight increases substantially to **0.34±0.04**, suggesting that static, baseline health indicators become more dominant in predicting limited improvement.

Similarly, when the model recommends "TR Patients," the **Video Keypoints** modality receives the highest attention (**0.28±0.05**), potentially reflecting the importance of visual feedback and observable progress in remote rehabilitation settings. For "CR Patients," **Clinical Data** again dominates (**0.33±0.05**), possibly indicating that patients who benefit more from conventional, in-person therapy might have more complex or established clinical profiles that influence treatment choice. This dynamic weighting mechanism validates the adaptive nature of Causal-MMFNet, allowing it to judiciously combine multimodal information based on the specific predictive context and patient characteristics.

4.10. Uncertainty Estimation Reliability

The integration of Monte Carlo Dropout (MC Dropout) for uncertainty estimation is crucial for providing clinicians with a measure of confidence alongside treatment recommendations. To evaluate the reliability of these uncertainty estimates, we assessed how well the predicted uncertainty (quantified as the standard deviation of MC Dropout predictions) correlates with the actual magnitude of prediction errors. A well-calibrated uncertainty mechanism should ideally assign higher uncertainty to predictions that are further from the true outcome.

Table 4 presents the average prediction errors (MAE) across different quantiles of predicted uncertainty. We divided the test set predictions into five bins based on their predicted uncertainty (from lowest to highest).

Table 4. Average Mean Absolute Error (MAE) across predicted uncertainty quantiles (Mean±Standard Deviation).

Uncertainty Quantile	Average Predicted Uncertainty	Average Prediction MAE
Lowest 20%	0.15±0.02	**1.88±0.04**
20% - 40%	0.25±0.03	2.15±0.05
40% - 60%	0.35±0.04	2.30±0.05
60% - 80%	0.45±0.04	2.58±0.06
Highest 20%	**0.60±0.05**	**2.89±0.07**

The results clearly demonstrate a positive correlation between the predicted uncertainty and the actual prediction error. Predictions falling into the lowest 20% uncertainty quantile exhibit an average MAE of **1.88±0.04**, indicating high accuracy where the model is most confident. Conversely, for predictions in the highest 20% uncertainty quantile, the average MAE rises significantly to **2.89±0.07**. This trend indicates that when Causal-MMFNet expresses higher uncertainty in its predictions, those predictions indeed tend to have larger errors. This validates the reliability of MC Dropout as an uncertainty quantification method within our framework, providing clinicians with a meaningful signal to gauge the trustworthiness of personalized recommendations and to identify cases that may warrant further clinical scrutiny or data collection.

5. Conclusion

In this study, we introduced Causal-MMFNet, a novel deep learning framework designed to provide personalized, data-driven decision support for stroke rehabilitation, specifically for opti-

 doi:10.20944/preprints202601.0873.v1

mizing TeleRehabilitation (TR) and Conventional Rehabilitation (CR) allocation. Leveraging diverse multimodal time-series data (IMU, video, logs, clinical data), Causal-MMFNet features a dynamic cross-modal attention mechanism and an Individual Treatment Effect (ITE) estimation module with causal consistency regularization. This architecture enables accurate prediction of balance function recovery and robust estimation of individualized treatment benefits, with Monte Carlo Dropout providing crucial uncertainty estimates. Our extensive evaluation on the *StrokeBalance-Sim* dataset demonstrated Causal-MMFNet's superior performance over baselines in both prediction and responder classification. Crucially, it significantly improved the quality of treatment allocation, leading to substantially greater actual BBS improvements for recommended patients. The framework represents a significant advancement towards precision rehabilitation, empowering clinicians with reliable, data-driven recommendations to optimize strategies, enhance patient outcomes, and improve resource utilization in stroke care.

References

1. Weijin Xu, Zhuang Sha, Huihua Yang, Rongcai Jiang, Zhanying Li, Wentao Liu, and Ruisheng Su. An automatic cascaded model for hemorrhagic stroke segmentation and hemorrhagic volume estimation. *CoRR*, 2024.
2. Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811. Association for Computational Linguistics, 2021.
3. Hao Wu, Hui Li, and Yiyun Su. Bridging the perception-cognition gap:re-engineering sam2 with hilbert-mamba for robust vlm-based medical diagnosis, 2025.
4. Yucheng Zhou, Lingran Song, and Jianbing Shen. Improving medical large vision-language models with abnormal-aware feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12994–13011, Vienna, Austria, July 2025. Association for Computational Linguistics.
5. Haimei Zhao, Jing Zhang, Zhuo Chen, Bo Yuan, and Dacheng Tao. On robust cross-view consistency in self-supervised monocular depth estimation. *Machine Intelligence Research*, 21(3):495–513, 2024.
6. Zhuo Chen, Haimei Zhao, Xiaoshuai Hao, Bo Yuan, and Xiu Li. Stvit+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization. *Applied Intelligence*, 55(5):328, 2025.
7. Xinjin Li, Yu Ma, Kaisen Ye, Jinghan Cao, Minghao Zhou, and Yeyang Zhou. Hy-facial: Hybrid feature extraction by dimensionality reduction methods for enhanced facial expression classification. *arXiv preprint arXiv:2509.26614*, 2025.
8. Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *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)*, pages 6153–6166. Association for Computational Linguistics, 2021.
9. Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767. Association for Computational Linguistics, 2023.
10. Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15890–15902. Association for Computational Linguistics, 2024.
11. Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*, 2024.
12. Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11964–11974, 2024.

 doi:10.20944/preprints202601.0873.v1

13. Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3008–3018, 2025.
14. Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021. Association for Computational Linguistics, 2021.
15. Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. Multimodal sentiment detection based on multi-channel graph neural networks. In *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)*, pages 328–339. Association for Computational Linguistics, 2021.
16. Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968. Association for Computational Linguistics, 2022.
17. Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507. Association for Computational Linguistics, 2023.
18. Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953. Association for Computational Linguistics, 2021.
19. Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959. Association for Computational Linguistics, 2021.
20. Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695. Association for Computational Linguistics, 2021.
21. Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172. Association for Computational Linguistics, 2021.
22. Luqing Ren et al. Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance. *Academic Journal of Computing & Information Science*, 8(8):8–14, 2025.
23. Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. PP-rec: News recommendation with personalized user interest and time-aware news popularity. In *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)*, pages 5457–5467. Association for Computational Linguistics, 2021.
24. Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. MINER: Multi-interest matching network for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 343–352. Association for Computational Linguistics, 2022.
25. Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. LLM-rec: Personalized recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612. Association for Computational Linguistics, 2024.
26. Yucheng Zhou, Jianbing Shen, and Yu Cheng. Weak to strong generalization for large language models with multi-capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.
27. Luqing Ren. Leveraging large language models for anomaly event early warning in financial systems. *European Journal of AI, Computing & Informatics*, 1(3):69–76, 2025.
28. Luqing Ren. Ai-powered financial insights: Using large language models to improve government decision-making and policy execution. *Journal of Industrial Engineering and Applied Science*, 3(5):21–26, 2025.

 doi:10.20944/preprints202601.0873.v1

29. Zhen Tian, Zhihao Lin, Dezong Zhao, Wenjing Zhao, David Flynn, Shuja Ansari, and Chongfeng Wei. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886*, 2025.
30. Haimei Zhao, Jing Zhang, Zhuo Chen, Shanshan Zhao, and Dacheng Tao. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14781–14791, 2024.
31. Liancheng Zheng, Zhen Tian, Yangfan He, Shuo Liu, Huilin Chen, Fujiang Yuan, and Yanhong Peng. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981*, 2025.
32. Zhihao Lin, Zhen Tian, Jianglin Lan, Dezong Zhao, and Chongfeng Wei. Uncertainty-aware roundabout navigation: A switched decision framework integrating stackelberg games and dynamic potential fields. *IEEE Transactions on Vehicular Technology*, pages 1–13, 2025.
33. Jingwei Yi, Fangzhao Wu, Chuhan Wu, Ruixuan Liu, Guangzhong Sun, and Xing Xie. Efficient-FedRec: Efficient federated learning framework for privacy-preserving news recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2824. Association for Computational Linguistics, 2021.