Towards Relaxed Multimodal Inputs for Gait-based Parkinson's Disease Assessment

Minlin Zeng, Zhipeng Zhou, Yang Qiu, Martin J. McKeown, Zhiqi Shen

***Abstract*—Parkinson's disease assessment has garnered growing interest in recent years, particularly with the advent of sensor data and machine learning techniques. Among these, multimodal approaches have demonstrated strong performance by effectively integrating complementary information from various data sources. However, two major limitations hinder their practical application: (1) the need to synchronize all modalities during training, and (2) the dependence on all modalities during inference. To address these issues, we propose the first Parkinson's assessment system that formulates multimodal learning as a multi-objective optimization (MOO) problem. This not only allows for more flexible modality requirements during both training and inference, but also handles modality collapse issue during multimodal information fusion. In addition, to mitigate the imbalance within individual modalities, we introduce a margin-based class rebalancing strategy to enhance category learning. We conduct extensive experiments on three public datasets under both synchronous and asynchronous settings. The results show that our framework—Towards Relaxed InPuts (`TRIP`)—achieves state-of-the-art performance, outperforming the best baselines by 16.48, 6.89, and 11.55 percentage points in the asynchronous setting, and by 4.86 and 2.30 percentage points in the synchronous setting, highlighting its effectiveness and adaptability.**

***Index Terms*—Parkinson's disease assessment, multi-objective optimization, multimodal information fusion**

I. Introduction

PARKINSON'S disease (PD) affects approximately 1–2% of the global population over the age of 65 and is second only to Alzheimer's disease among neurodegenerative disorders [1]. PD symptoms predominantly include tremor, speech impairment, and gait disturbances, arising from motor neuron degeneration. Symptom severity increases with disease progression; therefore, timely intervention is crucial to prevent deterioration and preserve quality of life. At present, PD assessment is mainly conducted by physicians using well-established clinical scales such as the Movement Disorder Society–Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Hoehn and Yahr (H&Y) scale [2]. However, in the absence of a definitive biomarker, the questionnaire-based assessment is time- and resource-intensive, subjective, and insufficiently sensitive to minor fluctuations [3]. Thus, there is a substantial need to streamline PD detection and progression assessment. With the growing availability of sensing technologies, researchers can now collect high-precision motion data from PD patients, providing rich, objective information that can complement traditional in-clinic evaluation.

Minlin Zeng, Zhipeng Zhou, Yang Qiu, and Zhiqi Shen are with College of Computing and Data Science, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (email: minlin001, webank-zpzhou, qiuyang, zqshen@e.ntu.edu.sg).

M. J. McKeown is with the Department of Medicine (Neurology), Pacific Parkinsons Research Centre, The University of British Columbia, Vancouver, BC V6T 2B5, Canada (email: martin.mckeown@ubc.ca).

TABLE I: Comparison of requirements of mainstream multimodal-based PD assessment approaches. 'Single' represents single modality-based approaches, while 'Early/Late Fusion', 'Cross Att.', and 'Shared Latent' represent the corresponding multimodal solutions.

	Single	Early/Late Fusion	Cross Att.	Shared Latent	`TRIP`
Multimodal	✗	✓	✓	✓	✓
Asynchronous	-	✗	✗	✗	✓
Optional Modality	✗	✗	✗	✓	✓

Many types of sensor data have been explored for the assessment of PD, including speech [4], movement [5], neuroimaging [6], and others. Among these modalities, gait-based movement data offer several advantages for diagnosis and severity monitoring [14]. First, gait disturbance is a prominent motor symptom even in the early stages of PD [8]. Second, quantitative gait analysis—explicitly endorsed in clinical guidelines (e.g., Timed Up and Go)—provides objective and reproducible metrics. Moreover, portable devices such as pressure sensors [9], depth sensors [10], cameras [11], their combinations [12], et al [13] can capture multi-perspective high-precision gait data, enabling a cost-effective yet multidimensional view of the pathophysiology of PD.

On the other hand, artificial intelligence (AI) algorithms, in particular, have been extensively applied in the existing literature to recognize disease-specific patterns stored in these gait data to aid in the evaluation of PD [14]. The current literature can be categorized into single- and multi-modality ones, where single-modality approaches rely on a single data source, while multimodal approaches integrate multiple data streams (e.g., combining camera and inertial sensors [10]) for a more comprehensive assessment. Although single-modality methods have shown competitive performance and currently dominate the field, an increasing body of work has demonstrated the advantages of multi-modality data in PD analytics [3], contributing to the growing popularity of multimodal solutions. Consequently, in this work we focus on multimodal gait data for PD assessment.

However, existing multimodal solutions rely on a strict data collection process, requiring synchronized inputs (i.e., strictly time-aligned signals) from full modalities during both training and inference (see Table I). This requirement presents several key challenges, including technical and personnel-related barriers. Technically, capturing synchronized multimodal gait data (e.g., vertical ground reaction force (vGRF) and skeleton data) often demands specialized equipment and complex,

TABLE II: Ablation accuracies (%) under synchronous input conditions. The two datasets, WearGait and FOG, consist of two classes with three modalities and three classes with two modalities, respectively. The modality and its results indicating modality collapse are highlighted in grey.

Dataset	Modality	Model 1	Model 2	Model 3
		Early Fusion	Late Fusion	Cross Attention
WearGait Dataset	w/o Insole	60.02	53.41	56.39
	w/o Walkway	61.72	56.84	52.01
	w/o IMU	**50.72**	**52.38**	**50.49**
	Full Modality	82.44	71.36	77.20
		Early Fusion	Late Fusion	Shared Latent
FOG Dataset	**w/o Sensor**	**36.42**	**36.71**	**36.14**
	w/o Skeleton	40.09	47.64	43.56
	Full Modality	61.65	63.71	61.13

non-standardized experimental setups. Personnel-related issues include data labeling and calibration for synchronizing input data [15]. Moreover, collecting all modalities during inference is often impractical due to privacy concerns and device limitations [16]. Therefore, an approach that handles missing modality and asynchronous streams can greatly alleviate these constraints.

In addition, during our experiments with some multimodal fusion models we observed a recurring failure mode which causes the fused model heavily leans on one subset of modalities and if this subset of modalities are removed during inference, the model's performance significantly degraded (see Table. II). This issue has serious effect on our scenario as missing modality is very common in real-clinic. In the synchronous setting, ablations make this asymmetry explicit: on WearGait, Early Fusion falls from 82.44% (full) to 50.72% when IMU is removed, yet remains 60.02%/61.72% after removing Insole/Walkway; on FOG, Early/Late/Shared drop to 36.42/36.71/36.14% without Sensor (near 33% chance for three classes), but stay at 40.09/47.64/43.56% without Skeleton. To the best of our knowledge, this phenomenon has not been discussed in gait-based PD assessment. Nevertheless, a subsequent review showed it is not unique to our field and has been discussed in multimodal computer vision field as "*modality collapse*", where fusion models rely on a subset of modalities or suppress one branch during optimization [17].

To address above mentioned challenges encountered frequently in real-clinic scenarios, we propose a modality-relaxed multimodal framework that allows asynchronous and optional modality inputs during both training and inference. As a first step, we design a new architecture (see Fig. 1) which enables interaction between modalities while preserving modality-specific features. Next, we employ a multi-objective optimization (MOO) algorithm that balances convergence and conflict avoidance to not only facilitate the learning of shared representations across modalities and subjects, but also mitigating the modality collapse issue. In addition, we introduce a class-rebalanced training scheme to mitigate class imbalance within each modality. Extensive experiments are conducted under both synchronous and asynchronous conditions. The experiment results show that our model not only achieves good accuracy under asynchronous condition, but also improves single modality accuracy compares to their single-modal baselines. In a nutshell, our contributions can be summarized as three-fold:

- To the best of our knowledge, we are the first to propose a practical PD assessment system that accommodates asynchronous modality inputs during training and allows optional modality inputs during inference.
- A MOO framework is developed from both architectural and optimization perspective to learn modality-shared features and help alleviate the modality collapse issue. In addition, a margin-based rebalancing strategy is proposed to promote balanced learning within each modality.
- Extensive experiments on three public datasets and two different input modes demonstrate that `TRIP` not only surpasses single-modality baselines but also outperforms conventional and more advanced multi-modal fusion strategies, highlighting the effectiveness and flexibility of our approach.

II. Related Work

Our work primarily lies at the intersection of PD assessment and MOO. In the following, we introduce key classical and recent developments in both domains and then explicitly highlight the connections and differences between our approach and existing research.

A. Gait-Based Single-Modality Solutions

Advancements in sensor devices and AI have enabled the shift in PD assessment from in-clinic evaluations with specialist supervision to AI-enhanced, automated, or remote monitoring approaches [18]–[20]. A common direction is to analyze gait data during locomotion, which includes kinematic data (e.g., RGB-D video [10], IMU [21]) and kinetic data (e.g., vGRF [22]). Early ML-based approaches use hand-crafted features (e.g., stride, speed, and cadence) with algorithms such as SVM [23], k-NN [24], or ensembles [7], but require extensive manual feature engineering. To reduce this reliance, DL models are introduced to learn representations directly from raw data [25], ranging from MLPs [26] to CNNs [27], LSTMs [28], and GNNs [29], each suited for spatial or temporal signals. More recent methods adopt transformers [30] and hybrid models [22] to capture complex spatiotemporal patterns. Overall, DL has been shown to consistently outperform ML in many PD tasks [31].

B. Gait-Based Multi-Modality Solutions

In addition to single-modality studies, multimodal approaches have demonstrated superior capacity in capturing holistic disease patterns by integrating diverse data sources [3]. However, effectively utilizing cross-modal information remains challenging. Early works either relied on ML with hand-crafted features [32] or combined modality-specific architectures (e.g., 3D CNNs and LSTMs [33], [34]), to extract spatial and temporal features separately. While these methods capture partial spatiotemporal or multimodal characteristics, they struggle with cross-modal correlation. More recent efforts can be categorized into three strategies (see Table I): early/late fusion, cross-attention, and shared latent representation. For instance, [10] proposed a late fusion method that uses separate spatial encoders followed by Correlative Memory Neural Networks to learn joint temporal embeddings. The work of [35] introduced a graph-based shared latent fusion by modeling each modality as graph vertices to learn inter-modality relationships. Another work [36] adopted cross-attention modules, applying bi-directional co-attention across silhouette and skeleton streams to extract complementary information.

Despite their effectiveness, these models face practical limitations. Most require either strict spatial-temporal alignment or complete modality availability during inference—constraints that are difficult to meet due to the complexity and cost of collecting fully synchronized multimodal gait data.

C. Multi-Objective Optimization

Multi-objective optimization (MOO) has been proposed to address machine learning problems involving multiple objectives. It can generally be categorized into two types: 1. Pareto Front Learning (PFL) and 2. Balanced Trade-off Exploration. PFL aims to approximate the entire Pareto front so that the desired trade-off can be achieved once user preferences are specified. In this direction, several methods have been introduced in recent years. PHN [37] uses a hypernetwork to generate Pareto-optimal models conditioned on preference vectors. COSMOS [38] reduces the parameter overhead by conditioning in the feature space rather than using a hypernetwork. PaMaL [39] learns the Pareto front in the manifold space by optimizing task-specific endpoints. To improve parameter efficiency, follow-up works have proposed low-rank approximations [40], [41] and mixture-of-experts (MoE) architectures [42] to reduce endpoint overhead.

On the other hand, balanced trade-off exploration is commonly used in multi-task learning (MTL), which seeks to achieve balanced progress across tasks. Various gradient-based MOO methods have been proposed in this context. MGDA [43] applies the Frank–Wolfe algorithm to find the gradient combination with the smallest norm. PCGrad [44] mitigates gradient conflicts by projecting gradients onto orthogonal subspaces. CAGrad [45] balances convergence and conflict avoidance through compromise objectives. Nash-MTL [46] adopts a game-theoretic approach, allowing tasks to negotiate parameter updates. Building on this, FairGrad [47] introduces a finer-grained constraint to ensure equitable learning progress among tasks.

Connection and Difference: Similar to prior work, our approach also leverages multimodal information to enhance PD assessment. However, as summarized in Table I, a key distinction is that our proposed MOO framework eliminates the requirement for synchronous multimodal inputs during both training and inference, thereby improving practicality and aligning better with real-world deployment scenarios. To the best of our knowledge, the most related work is that of Heidarivincheh et al. [48], which employs a variational autoencoder (VAE) to encode all modalities into a shared latent space, allowing for optional modality input at inference. Nevertheless, their method does not support asynchronous multimodal input during training, a limitation that our work explicitly addresses.

III. Principal Design

In this section, we present a detailed overview of our proposed framework (see Fig. 1), architectural design, MOO learning paradigm, and class rebalancing strategy.

A. Problem Setup

We consider multimodal gait-based PD assessment with m input modalities per subject. Let

$$\mathcal{D} = \big\{(x_1^i, x_2^i, \dots, x_m^i,\, y^i)\big\}_{i=1}^{N},$$

where $x_r^i \in \mathbb{R}^{T_r \times D_r}$ denotes the sequence window from modality $r \in \{1, \dots, m\}$ and $y^i \in \{1, \dots, K\}$ is the clinical label. At training and inference, a subset of modalities $\mathcal{S}_i \subseteq \{1, \dots, m\}$ may be available due to asynchrony or dropout. A model f_Θ consumes any available subset and predicts

$$f_\Theta\big(\{x_r^i\}_{r \in \mathcal{S}_i}\big) \;\rightarrow\; \hat{y}^i.$$

B. Overall Framework

For each modality r, a modality-specific encoder $e_r(\cdot;\,\omega_r)$ maps the raw sequence to a fixed-width feature sequence

$$u_r^i = e_r\big(x_r^i;\, \omega_r\big).$$

Each u_r^i is then processed by a *shared* encoder $g(\cdot;\,\varphi)$ to produce shared representations

$$s_r^i = g\big(u_r^i;\, \varphi\big).$$

A modality-specific head $h_r(\cdot;\,\theta_r)$ produces logits

$$z_r^i = h_r\big(s_r^i;\, \theta_r\big),$$

and a per-modality supervised loss $\ell_r^i = \ell\big(z_r^i, y^i\big)$ is computed for every $r \in \mathcal{S}_i$. Private parameters $\{\omega_r, \theta_r\}$ are updated using their respective gradients from ℓ_r. Gradients on the shared encoder φ induced by $\{\ell_r\}_{r \in \mathcal{S}_i}$ are combined by a MOO step to yield a single update for φ. At inference, any subset of modalities can be provided.

C. Architecture Design

We adopt a modular *encoders* $\rightarrow$ *shared backbone* $\rightarrow$ *task heads* design. It supports relaxed inputs at test time and cleanly separates private vs. shared parameters for multi–objective training. (1) *Encoders*. Each modality uses a modality-specific encoder to convert raw signals into embeddings with a common feature width to interface with the same backbone. (2) *Shared Backbone*. A single backbone is reused across all streams, which transforms each encoder's embedding into a compact representation for classification. The same weights are shared across modalities to encourage cross–modal regularization. (3) *Task Heads.* Backbone features are flattened and routed to classification heads. Synchronous inputs use one shared head referenced by all streams; asynchronous inputs

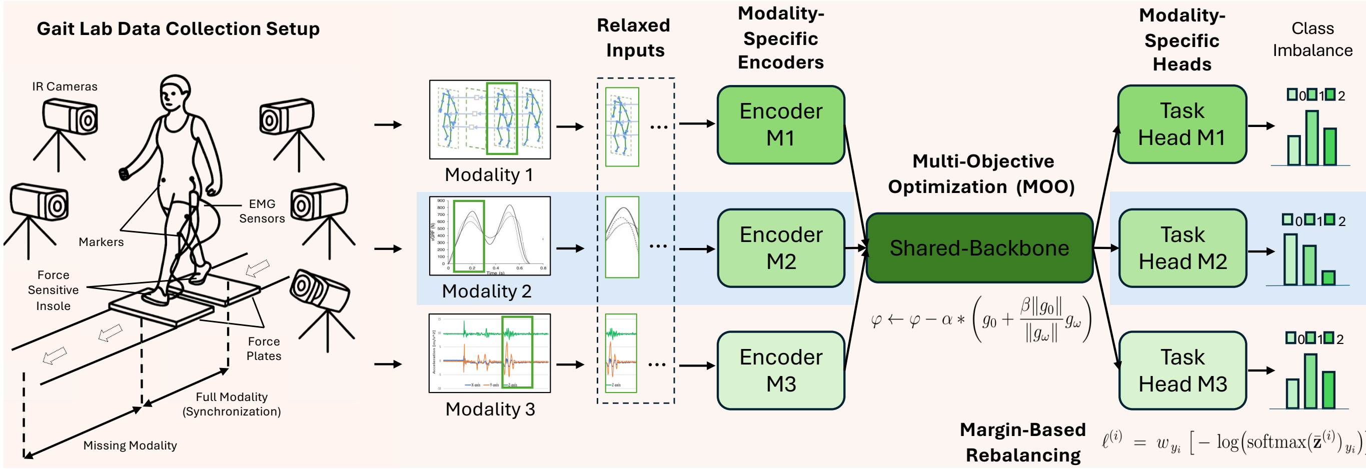

Fig. 1: `TRIP` overview (relaxed-input scenario). Gait lab setup: conventional labs record multiple synchronized modalities (e.g., 3D motion capture and wearables), but trials often have missing streams or dropouts. Our pipeline supports asynchronous inputs and missing modalities through: (1) **Relaxed Inputs**: trains and infers with time-misaligned clips and enabled missing modalities (blue bands); each stream is encoded independently, then processed by a shared backbone and modality heads—matching real deployment. (2) **Shared-Only MOO**: modality losses update their own private parameters, while only the shared parameters φ receive mixed, conflict-averse gradients using our proposed MOO objective. (3) **Imbalance-Aware Heads**: our margin-based rebalancing strategy at each head reweights classes with cosine-normalized logits to handle the inter/intra-class imbalance typical of multimodal gait PD data.

use one head per modality to permit training and inference with any subset of available data. Based on training conditions, private parameters are the encoders and per–modality heads in asynchronous condition; shared parameters are the backbone and the shared head in synchronous condition.

However, while this design allows flexible inputs at inference and clearly separates modality-specific (private) parameters, conflicts can arise during training on the shared parameters (see Fig. 2, left). Therefore, a strategy that resolves these potential conflicts is crucial for stable training and smooth gradient updates.

D. MOO-Based Multimodal Learning

To further enable the asynchronous and optional modality input, we develop a MOO learning paradigm to facilitate the shared feature extraction on shared backbone (f_φ) across modalities.

Assume the losses from all modality are $\{\mathcal{L}_i\}_{i=1}^M$ and their derived gradients on φ are $\{g_i\}_{i=1}^M$, where M is the number of modality (set as 2 or 3 in this paper). Since our objective is to promote all modalities' learning rather than optimizing the average ones, a worst case optimization is adopted as follow:

$$\min_d \max_i \frac{1}{\alpha}(\mathcal{L}_i(\varphi - \alpha * d) - \mathcal{L}_i(\varphi)) \tag{1}$$

where d is the update vector for φ, i.e., $\varphi \leftarrow \varphi - \alpha * d$, and α is the learning rate. This formulation first seeks the least progress one, and then optimize d to improve it. Note that Eqn. 1 can be further transformed as follow by applying first-order Taylor approximation:

$$\min_d \max_i - g_i^\top d \tag{2}$$

where g_i is $\nabla\mathcal{L}_i(\varphi)$. Note that $\max_i - g_i^\top d = \max_{\omega\in[M]} - (\sum_i \omega_i g_i)^\top d$, where $[M]$ is a simplex. Then we impose a norm constraint on d in Eqn. 2, i.e., $\|d\| - \|g_0\| \leq \|d - g_0\| \leq \beta\|g_0\|$. g_0 is the average gradient of all modalities, which serves as a naïve optimization. $\beta \in [0, 1)$ is a hyper-parameter measures the allowed deviation from g_0. This constraint ensures the convergence of MOO algorithm. Therefore, we formulate the dual objective according to Lagrangian equation and Slater's condition as follow:

$$\min_{\omega\in[M]} \max_d g_\omega^\top d - \lambda(\|d - g_0\|^2 - \beta\|g_0\|^2), \qquad \lambda > 0 \tag{3}$$

where $g_\omega = \sum_i \omega_i g_i$. By fixing λ and ω, we can obtain the optimal $d = g_0 + g_\omega/\lambda$. Inserting the optimal d into Eqn. 3, we have:

$$\min_{\omega\in[M],\ \lambda} g_\omega^\top g_0 + \frac{1}{2\lambda}\|g_\omega\|^2 + \frac{\lambda}{2}\beta^2\|g_0\|^2 \tag{4}$$

By applying the mean value inequality, we have the final objective as follow (when $\lambda = \frac{\|g_\omega\|}{\beta\|g_0\|}$):

$$\min_{\omega\in[M]} g_\omega^\top g_0 + \beta\|g_0\|\|g_\omega\| \tag{5}$$

Once the optimal ω^* are derived according to Eqn. 5, we have $d = g_0 + \frac{\beta\|g_0\|}{\|g_\omega\|} g_\omega$, and φ is updated as follow:

$$\varphi \leftarrow \varphi - \alpha * \left(g_0 + \frac{\beta\|g_0\|}{\|g_\omega\|} g_\omega\right) \tag{6}$$

This process is illustrated in Fig. 2. The employed MOO algorithm adjusts the weighting of modality-specific gradients to promote conflict-averse and balanced progress across individual tasks.

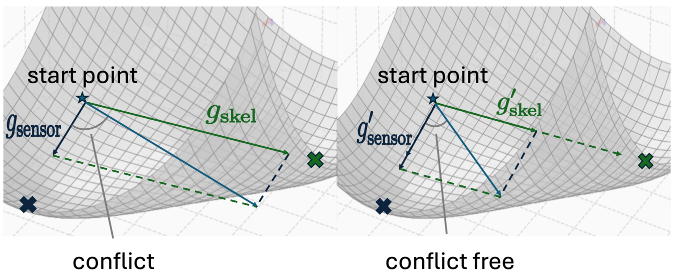

Fig. 2: Illustration of the MOO learning paradigm (two-modalities scenarios). Left: gradient conflict between the update direction of the sensor modality (g_{sensor}) and the overall average gradient direction. Right: conflict resolved after applying MOO optimization, aligning the adjusted modality gradients (g'_{sensor}, g_{skel}) with the shared descent direction.

Remarks: Only shared backbone's parameters φ are updated via Eqn. 6, other modules are updated by their corresponding modality-specific gradients (see Algorithm 1).

E. Margin-Based Rebalancing Strategy

Another issue lies in the prevalent class imbalance in PD multimodal datasets. As depicted in Fig. 3, both FOG and FBG exhibit a long-tailed distribution across categories. To mitigate the adverse effects of long-tailed class distribution inherent in PD multimodal gait classification tasks, we integrate class-adaptive margins and stochastic smoothing into the standard cross-entropy framework for both modalities.

Concretely, let K denote the total number of classes and let N_j be the number of training samples for class j. We first construct a *per-class margin vector* $\mathbf{m} = (m_1, \dots, m_C)^\top$ whose entries are proportional to $\log(N_{\max}) - \log(N_j)$, where $N_{\max} = \max_k N_k$. That is,

$$m_j \;=\; \log\bigl(N_{\max}\bigr) \;-\; \log\bigl(N_j\bigr), \qquad j = 1, \dots, K. \tag{7}$$

During the forward pass, suppose a given batch of size B produces cosine-normalized logits $\mathbf{z}^{(i)} = \bigl(z_1^{(i)}, \dots, z_K^{(i)}\bigr)$ for (i=1,...,B). We then inject a small, class-scaled random perturbation $\delta_j^{(i)} \sim \mathcal{N}(0, \sigma^2)$ (clamped to $[-1, 1]$) into each logit according to

$$\tilde{z}_j^{(i)} \;=\; z_j^{(i)} \;-\; \eta\,\bigl|\delta_j^{(i)}\bigr|\,\frac{m_j}{\max_k m_k}, \tag{8}$$

where $\eta > 0$ is a noise-magnitude hyperparameter that ensures more perturbation for under-represented classes. Next, we apply an *additive margin* m only to the target-class entry: if y_i is the ground-truth label for sample i, then

$$\hat{z}_j^{(i)} \;=\; \begin{cases} \tilde{z}_{y_i}^{(i)} \;-\; m, & j \;=\; y_i, \\ \tilde{z}_j^{(i)}, & j \;\neq\; y_i. \end{cases} \tag{9}$$

In addition, we incorporate *logarithmically scaled class weights* to compensate for imbalance. Define

$$w_j \;=\; \frac{\log\bigl(N_{\max}/N_j + \varepsilon\bigr)}{\text{div}}, \qquad j = 1, \dots, K, \tag{10}$$

where $\varepsilon > 0$ prevents $\log(0)$ and div is a tunable divisor that controls the overall weight magnitude. We then normalize $\mathbf{w} = (w_1, \dots, w_C)$ so that $\sum_{j=1}^{C} w_j = C$. The final loss for sample i is

$$\ell^{(i)} \;=\; w_{y_i}\,\bigl[-\,\log\bigl(\operatorname{softmax}(\bar{\mathbf{z}}^{(i)})_{y_i}\bigr)\bigr]. \tag{11}$$

Averaging $\ell^{(i)}$ over the batch yields the overall loss. This combination of *(i)* class-adaptive margin subtraction, *(ii)* logarithmic re-weighting, and *(iii)* noise smoothing directs the model's capacity toward underrepresented gait classes while maintaining stability on majority classes. The overall training scheme is summarized in Algorithm 1.

1) Integrated Loss Function:

Algorithm 1: Training Paradigm of `TRIP`

Input: Training Dataset $\mathcal{D} = \bigl\{(x_1^i, \dots, x_m^i,\, y^i)\bigr\}_{i=1}^{N}$
Output: Model trained with `TRIP`

Stage 1:
Initialize $\{\omega_r,\, \varphi,\, \theta_r\}$ randomly, with $r \in \mathcal{S}_i$ and $\mathcal{S}_i \subseteq \{1, \dots, m\}$
while not converged **do**
  foreach *batch* $\mathcal{B}_i$ *in* $\mathcal{D}$ **do**
    Compute modality-specific loss ℓ_r via Eqn. 11.
    Derived gradients on $\{\omega_r,\, \varphi,\, \theta_r\}$ with respect to ℓ_r.
    Update φ with g_φ^r via Eqn. 6.
    $\omega_r \leftarrow \omega_r - \alpha * g_{\omega_r}$,
    $\theta_r \leftarrow \theta_r - \alpha * g_{\theta_r}$.

IV. Implementation

We evaluate single-modality and multimodal fusion baselines on three gait datasets for PD, using subject-wise stratified cross-validation (CV). All models share the same classification setup for fair comparison.

A. Setup

a) Sampling Strategy: To mitigate imbalances, we use two complementary strategies. (1) *Modality-Balanced Training* (asynchronous only): in asynchronous training, the sample size of each modality varies, so we perform modality-level oversampling. However, this does not guarantee a class-balanced training set within each modality. To address this, we propose a margin-based class rebalancing strategy, which helps ensure class-balanced training. This stabilizes gradient contributions and enhances joint representation learning alongside the MOO objective. (2) *Class- and Modality-Balanced Evaluation* (synchronous and asynchronous): for fair validation, we oversample the evaluation set to equalize class counts and segment counts within and across modalities.

b) Training and Hyperparameters: Hyperparameter tuning is performed per dataset and model configuration, optimizing for classification accuracy. All models are trained using the Adam optimizer for a maximum of 100 epochs and the learning rate is fixed throughout. Experiments are conducted using an NVIDIA Tesla V100 GPU.

B. Training Strategy

We use two different training strategies in this work as detailed below. (1) **Synchronous**: this setting is commonly set as default in previous models which requires inputs to be time-aligned and from the same subject during training. (2) **Asynchronous**: this setting is designed to simulate real-world deployments where modality dropouts and time misalignment are common at inference. During training, inputs from different modalities may come from different subjects and from different time segments. At inference, when a modality is missing, the model processes the available modality. This increases usable data and diversity through relaxing strict alignment. It also reflects clinical reality (missing/noisy streams) and encourages subject-invariant representations by exposing the shared encoder to cross-subject variability during training, while evaluation remains subject-wise to avoid leakage.

Crucially, the asynchronous setting is ill-posed for the baseline models as mixing subjects introduces cross-label contamination. We report their asynchronous results (see Table III) only for completeness—to highlight this limitation and to compare against our design, which is expressly built for asynchronous inputs.

C. Validation Strategy

a) Cross Validation: We use subject-wise stratified k-fold CV. Subjects are grouped by class; each fold's evaluation set contains one subject per class (so k is bounded by the smallest class). For multimodal experiments, only subjects with complete data in both modalities are eligible for the evaluation set; single-modality baselines reuse the same evaluation subjects. Remaining subjects form the training set. This prevents leakage, preserves train/eval independence, and yields balanced estimates.

b) Baselines: We evaluate on both single modality and multimodal fusion baselines. All models inherit similar design of encoders, backbone, and training hyper-parameters as our pipeline to isolates fusion effects from all other factors under both synchronous and asynchronous settings. **Single-Modality Baselines**: They use the exact architecture from our main pipeline, but only one encoder is active at a time. These serve as upper bounds and isolate modality–specific signal. **Conventional Methods**: They serve as naive fusion baselines for our approach. (1) *Early/Late Fusion*: Early Fusion concatenates low-level features from encoders and feeds the joint sequence to the shared backbone; Late Fusion processes each modality independently through the backbone, then concatenates the resulting high-level vectors for prediction. (2) *Shared-Latent Fusion*: Each modality is linearly projected into a common latent width and fused by element-wise addition before the backbone. (3) *Cross-Attention Fusion*: A lightweight, symmetric cross-attention block lets each modality attend to the other; the fused sequence is then passed to the backbone.

SOTA Variants: These are implemented to benchmark our method against newer designs. (1) *FOCAL* [49]: This method factorizes each modality into shared and private latents, enforces orthogonality between them, and applies a contrastive objective to align shared parts across modalities, preserving complementary cues while reducing single-stream shortcuts. (2) *TACA* [50]: It compresses sequences into a small set of learned tokens per modality and performs cross-attention among tokens (plus a time-shared module) for efficient, salient fusion that tolerates time-misaligned streams. (3) *DEEPAV* [51]: This method use lightweight per-modality streams followed by late cross-modal interaction and an agreement/consensus head, emphasizing decision-level fusion and robust fallback when one modality is weak or missing.

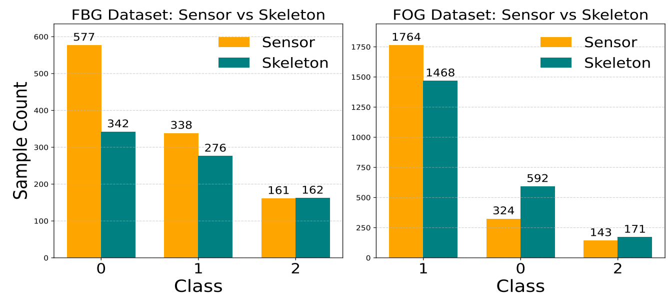

Fig. 3: Class distribution of multimodal PD datasets (FBG and FOG). Both datasets exhibit long-tailed distributions and noticeable intra-class imbalance between sensor and skeleton modalities.

V. Evaluation

A. Evaluation Metric

In addition to reporting individual performance and mean accuracy, we also incorporate a widely used metric, $\Delta m\%$ [55], which evaluates the overall degradation compared to independently trained models that are considered as the reference oracles. The formal definition of $\Delta m\%$ is given as:

$$\Delta m\% = \frac{1}{M}\sum_{i=1}^{M}(-1)^{\delta_i}(P_{m,i} - P_{b,i})/P_{b,i} \qquad (12)$$

where $P_{m,i}$ and $P_{b,i}$ represent the metric P_i for the compared method and the single modality-based model, respectively. The value of δ_i is assigned as 1 if a higher value is better for P_i, and 0 otherwise.

B. Parkinson's Disease Multimodal Datasets

We conducted evaluations on three publicly available multimodal PD gait datasets. All datasets contain gait data from two to three complementary modalities and accompanied by clinically annotated ratings. These characteristics make them suitable for the evaluation of our proposed framework.

FOG Dataset [52]: It contains video and IMU recordings from 35 PD subjects performing turning-in-place at 30 Hz (video) and 128 Hz (IMU) with clinical FoG annotations. Coordinates of 3D poses are obtained using MMPose [56] (MotionBERT trained on Human3.6M). **FBG Dataset** [53]: It comprises synchronized full-body kinematics and kinetics from 26 PD subjects walking overground in ON/OFF medication states, including 3D motion capture at 150 Hz and force-plate data, with MDS-UPDRS III and H&Y labels. **WearGait**

TABLE III: Performance comparison across datasets (Accuracy ± Std) under asynchronous input conditions. Left block uses inverse-frequency class rebalancing (w/). Right block uses original training (w/o). Accuracy (mean±std) results are obtained on class-balanced test sets, where class balance ensures fair performance comparison without the need for F1-score reporting. The accuracy results of each single modality are calculated from each task-specific stream, while the average accuracy is simply the average of them. All models use similar capacity (≈10k params; similar backbone structure). The best result is highlighted in **bold**. The second-best result is underlined. Lower $\Delta m\%$ is better.

Dataset	Approach	**w/** class rebalancing					**w/o** class rebalancing				
		Mod.1	Mod.2	Mod.3	Avg Acc ↑	$\Delta m\%\downarrow$	Mod.1	Mod.2	Mod.3	Avg Acc ↑	$\Delta m\%\downarrow$
FOG Dataset		**Skeleton**	**IMU**	**—**			**Skeleton**	**IMU**	**—**		
	Skeleton Only	$\underline{58.74 \pm 8.65}$	–	–	–	–	$\underline{47.73 \pm 6.07}$	–	–	–	–
	Sensor Only	–	50.50 ± 3.45	–	–	–	–	33.47 ± 0.14	–	–	–
	Early Fusion	41.59 ± 3.53	53.41 ± 8.39	–	47.50 ± 2.96	11.72	41.39 ± 3.72	35.55 ± 2.28	–	38.47 ± 1.73	3.53
	Late Fusion	44.48 ± 4.50	50.70 ± 4.41	–	47.59 ± 2.50	12.21	44.16 ± 2.78	36.25 ± 1.72	–	$\underline{40.20 \pm 1.74}$	$\underline{-0.41}$
	Shared Latent	37.81 ± 2.40	53.35 ± 6.08	–	45.58 ± 3.07	14.99	40.94 ± 3.44	$\underline{36.66 \pm 2.43}$	–	38.80 ± 2.10	2.35
	Cross Attention	38.80 ± 3.43	48.33 ± 6.08	–	43.57 ± 3.11	19.12	38.99 ± 1.78	36.27 ± 2.15	–	37.63 ± 1.41	4.97
	FOCAL [49]	48.59 ± 5.09	$\mathbf{53.91 \pm 6.19}$	–	51.26 ± 3.88	5.26	42.66 ± 3.21	34.81 ± 1.74	–	38.73 ± 1.68	3.31
	TACA [50]	45.27 ± 3.02	39.02 ± 2.83	–	42.14 ± 2.43	22.83	39.07 ± 1.52	33.39 ± 0.16	–	36.23 ± 0.81	9.19
	DEEPAV [51]	53.69 ± 9.00	$\underline{53.47 \pm 1.26}$	–	$\underline{54.03 \pm 4.46}$	$\underline{1.36}$	45.35 ± 2.84	34.37 ± 1.64	–	39.86 ± 2.14	1.15
	`TRIP`	$\mathbf{63.61 \pm 3.86}$	49.76 ± 4.97	–	$\mathbf{56.68 \pm 4.17}$	$\mathbf{-3.41}$	$\mathbf{63.61 \pm 3.86}$	$\mathbf{49.76 \pm 4.97}$	–	$\mathbf{56.68 \pm 4.17}$	$\mathbf{-40.97}$
FOG Dataset		**Skeleton**	**vGRF**	**—**			**Skeleton**	**vGRF**	**—**		
	Skeleton Only	53.31 ± 4.26	–	–	–	–	41.80 ± 4.46	–	–	–	–
	Sensor Only	–	$\mathbf{68.78 \pm 3.04}$	–	–	–	–	60.13 ± 3.12	–	–	–
	Early Fusion	39.94 ± 3.35	62.50 ± 3.04	–	51.22 ± 1.38	17.11	35.27 ± 1.41	63.32 ± 4.37	–	49.30 ± 2.39	5.16
	Late Fusion	46.18 ± 4.03	61.70 ± 4.38	–	53.94 ± 0.64	11.83	41.52 ± 3.41	63.64 ± 2.81	–	$\underline{52.59 \pm 1.71}$	-2.58
	Shared Latent	40.39 ± 2.07	63.18 ± 2.33	–	51.79 ± 0.64	16.19	35.03 ± 1.66	$\underline{64.49 \pm 3.55}$	–	49.76 ± 1.91	4.47
	Cross Attention	34.51 ± 1.21	57.66 ± 3.79	–	46.08 ± 2.35	25.72	33.47 ± 0.15	$\underline{54.82 \pm 5.49}$	–	44.14 ± 2.75	14.38
	FOCAL [49]	54.08 ± 4.55	62.07 ± 3.71	–	58.07 ± 3.52	4.16	51.54 ± 4.59	50.12 ± 5.95	–	50.83 ± 4.04	-3.33
	TACA [50]	$\underline{56.32 \pm 1.87}$	49.70 ± 2.17	–	53.01 ± 1.32	11.04	$\underline{59.61 \pm 2.62}$	37.62 ± 0.99	–	48.62 ± 1.57	-2.59
	DEEPAV [51]	$\mathbf{63.58 \pm 6.32}$	52.74 ± 1.91	–	$\underline{58.16 \pm 3.10}$	**2.03**	$\mathbf{62.12 \pm 3.00}$	42.41 ± 3.03	–	52.27 ± 1.71	$\underline{-9.57}$
	`TRIP`	51.24 ± 5.76	$\underline{67.71 \pm 3.46}$	–	$\mathbf{59.48 \pm 3.03}$	$\underline{2.72}$	51.24 ± 5.76	$\mathbf{67.71 \pm 3.46}$	–	$\mathbf{59.48 \pm 3.03}$	$\mathbf{-17.59}$
WearGait Dataset		**Walkway**	**Insole**	**IMU**			**Walkway**	**Insole**	**IMU**		
	Walkway Only	66.22 ± 1.61	–	–	–	–	63.07 ± 2.94	–	–	–	–
	Insole Only	–	$\underline{59.55 \pm 1.39}$	–	–	–	–	$\underline{59.23 \pm 1.03}$	–	–	–
	IMU Only	–	–	$\underline{77.56 \pm 1.34}$	–	–	–	–	$\underline{77.30 \pm 1.41}$	–	–
	Early Fusion	54.31 ± 1.28	56.25 ± 1.42	55.19 ± 0.92	55.25 ± 0.92	17.46	52.58 ± 1.38	55.97 ± 0.82	53.55 ± 1.24	53.94 ± 3.66	17.62
	Late Fusion	62.01 ± 3.37	54.63 ± 1.60	65.32 ± 1.84	60.65 ± 1.57	10.13	54.64 ± 2.14	54.77 ± 1.80	62.69 ± 2.31	57.37 ± 0.95	13.27
	Shared Latent	59.98 ± 1.70	55.78 ± 1.55	59.25 ± 1.77	58.34 ± 0.70	13.12	51.58 ± 1.20	56.79 ± 1.51	54.38 ± 2.13	54.25 ± 1.02	17.33
	Cross Attention	53.24 ± 0.95	53.97 ± 1.57	52.90 ± 1.30	53.37 ± 0.91	20.26	52.42 ± 0.82	52.70 ± 1.63	52.60 ± 1.21	52.58 ± 0.79	19.95
	FOCAL [49]	60.60 ± 1.30	55.72 ± 1.71	61.33 ± 1.21	59.22 ± 0.95	11.95	50.96 ± 1.13	57.37 ± 1.33	58.18 ± 3.10	55.50 ± 1.10	15.69
	TACA [50]	$\underline{66.62 \pm 0.80}$	54.42 ± 2.01	61.21 ± 1.55	$\underline{60.75 \pm 0.59}$	$\underline{9.70}$	$\underline{65.32 \pm 1.25}$	54.43 ± 1.75	59.76 ± 1.36	$\underline{59.84 \pm 0.78}$	$\underline{9.08}$
	DEEPAV [51]	53.08 ± 2.53	56.91 ± 1.98	52.66 ± 1.52	54.21 ± 1.02	18.79	52.71 ± 0.63	57.74 ± 0.90	52.90 ± 1.36	54.45 ± 0.32	16.84
	`TRIP`	$\mathbf{71.08 \pm 1.86}$	$\mathbf{63.03 \pm 1.04}$	$\mathbf{80.07 \pm 1.77}$	$\mathbf{71.39 \pm 1.09}$	$\mathbf{-5.47}$	$\mathbf{71.08 \pm 1.86}$	$\mathbf{63.03 \pm 1.04}$	$\mathbf{80.07 \pm 1.77}$	$\mathbf{71.39 \pm 1.09}$	$\mathbf{-8.69}$

Dataset [54]: It provides synchronized IMU, sensorized insole (16-sensor plantar pressures and embedded inertial signals), and walkway data, all at 100Hz, from 98 people with PD and 83 age-matched controls with clinical metadata.

C. Overall Evaluation

We evaluate the proposed `TRIP` framework under both asynchronous and synchronous conditions, each with and without class re-balancing, to comprehensively assess robustness and generalization. Single-modality oracles (e.g., Skeleton Only) serve as upper bounds, while $\Delta m\%$ quantifies how multimodal fusion compares to its strongest unimodal counterpart (negative indicates improvement).

Asynchronous Inputs. Under naturally imbalanced, asynchronous inputs—typical in clinical deployment—`TRIP` consistently achieves the highest accuracies (right blocks in Table III). For FOG, `TRIP` attains 56.68%, surpassing the best baseline (Late Fusion, 40.2%) and yielding a large multimodal gain (Δm=-41.97). Skeleton accuracy rises from 47.73% to 63.61% (+15.88 pp), and IMU from 33.47% to 49.76% (+16.29 pp). For FBG, it reaches 59.48%, outperforming DEEPAV (52.59%) by +6.89 pp, with Skeleton improved from 41.8% to 51.24% (+9.44 pp) and vGRF from 60.13% to 67.71% (+7.58 pp), resulting in Δm=-17.59. For WearGait, which includes three modalities, `TRIP` achieves 71.39%, outperforming all fusion methods while maintaining strong per-modality accuracies (Walkway 71.08%, Insole 63.03%, IMU 80.07%). These results confirm that `TRIP` can effectively learn from time-misaligned data without synchronized sampling, while other methods deteriorate.

Synchronous Inputs. When all modalities are synchronized (Table IV), `TRIP` retains its lead (right blocks). On FOG, it achieves 62.35%, outperforming all fusion baselines and exceeding the best unimodal oracle (57.49%) by +4.86 pp (Δm=-9.46). On WearGait, `TRIP` reaches 84.18%, the best among all models, with a substantial multimodal gain (Δm=-28.17). It surpasses all fusion competitors (e.g., Early Fusion 81.88%, Cross Attention 79.3%) and all single-modality oracles (Walkway 64.01%, Insole 59.59%, IMU 76.6%). We omit FBG synchronous results since inconsistent temporal alignment across sensor trials prevents fair synchronized evaluation, further underscoring the limitations of sync-dependent approaches.

Effect of Class Rebalancing. To confirm that improvements

TABLE IV: Performance of FOG and WearGait datasets (Accuracy ± Std) under synchronous input condition. Left block uses inverse-frequency class rebalance (w/). Right block uses original training (w/o). The best result is highlighted in **bold**. The second-best result is underlined. Lower $\Delta m\%$ is better.

Dataset	Approach	w/ class rebalancing		w/o class rebalancing	
		Acc (%) ↑	$\Delta m\%$ ↓	Acc (%) ↑	$\Delta m\%$ ↓
FOG	Skeleton Only	56.95 ± 5.10	–	56.44 ± 7.27	–
	IMU Only	61.09 ± 4.20	–	57.49 ± 4.52	–
	Early Fusion	61.31 ± 6.15	-4.01	48.54 ± 5.55	14.78
	Late Fusion	62.32 ± 5.73	-5.72	50.11 ± 4.54	12.03
	Shared Latent	61.03 ± 7.20	-3.53	45.76 ± 3.90	19.66
	Cross Attention	55.86 ± 7.84	5.24	47.37 ± 4.13	16.84
	FOCAL [49]	56.47 ± 10.41	4.20	46.87 ± 4.46	17.71
	TACA [50]	60.05 ± 5.47	-1.87	46.27 ± 3.71	18.77
	DEEPAV [51]	61.35 ± 7.99	-4.08	51.55 ± 4.09	9.50
	TRIP	$\mathbf{62.35 \pm 4.38}$	$\mathbf{-5.77}$	$\mathbf{62.35 \pm 4.38}$	$\mathbf{-9.46}$
WearGait	Walkway Only	66.59 ± 1.63	–	64.01 ± 2.10	–
	Insole Only	59.77 ± 0.91	–	59.59 ± 1.11	–
	IMU Only	77.14 ± 2.23	–	76.60 ± 2.49	–
	Early Fusion	82.55 ± 1.64	-23.03	81.88 ± 6.06	-24.07
	Late Fusion	72.28 ± 1.69	-7.72	69.44 ± 2.65	-5.22
	Shared Latent	60.86 ± 2.20	9.30	56.53 ± 0.64	14.34
	Cross Attention	77.36 ± 0.97	-15.30	79.30 ± 1.58	-20.16
	FOCAL [49]	72.92 ± 0.83	-8.68	71.78 ± 1.58	-8.77
	TACA [50]	58.39 ± 0.78	12.98	74.04 ± 4.69	-12.19
	DEEPAV [51]	76.46 ± 1.61	-13.95	50.74 ± 3.14	23.11
	TRIP	$\mathbf{84.18 \pm 1.84}$	$\mathbf{-25.46}$	$\mathbf{84.18 \pm 1.84}$	$\mathbf{-28.17}$

are not caused by label imbalance, we repeat experiments with inverse-frequency weighting on the loss function for all baseline methods. Under *asynchronous + rebalanced* training (left blocks in Table III), TRIP maintains its superiority in mean accuracy: 56.68% on FOG (Δm=−3.41), 59.48% on FBG (Δm=2.72), and 71.39% on WearGait (Δm=−5.47). Under *synchronous + rebalanced* training (left blocks in Table IV), TRIP again performs best: 62.35% on FOG (Δm=−5.77) and 84.18% on WearGait (Δm=−25.46). Even when class frequencies are equalized, TRIP outperforms or matches the best rebalanced baselines (e.g., DEEPAV*, Early Fusion*), while preserving high per-modality accuracies.

Summary. Across all configurations—synchronous or asynchronous and balanced or unbalanced—TRIP consistently delivers the strongest fusion accuracy and the lowest $\Delta m\%$. These findings confirm that: (1) TRIP robustly integrates complementary modalities without temporal alignment; (2) it enhances weaker modalities while maintaining strong ones via gradient-balanced optimization; and (3) it remains stable under class rebalancing, demonstrating intrinsic robustness to both modality and label imbalance.

D. Modality Collapse Mitigation

Apart from effective multimodal information fusion, compared to all other fusion baselines, TRIP keeps much higher single and pairwise (i.e., combine two modalities) accuracies when we zero-mask other streams at test time. As shown in Fig. 4, most of other baselines' results hover around 50–60% under the same masks. In addition, pairwise drops from full are smaller, indicating graceful degradation rather than failure

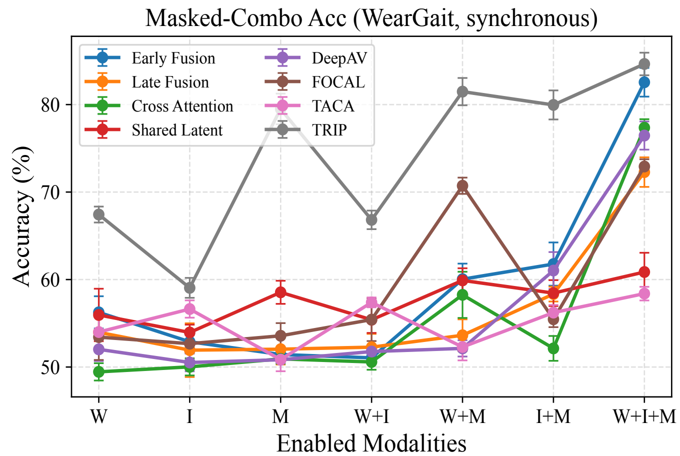

Fig. 4: Demonstration of effective modality collapse mitigation with TRIP. Results are obtained using WearGait dataset under synchronous inputs condition. Values in x-axis include all combinations of different modalities used during inference (W: Walkway, I: Insole, M: IMU).

when one modality is missing. That means TRIP learned useful per-stream representations instead of collapsing to one single modality.

TABLE V: Ablation study on important modules/loss functions on three datasets (Accuracy ± Std) with asynchronous inputs. The best result is highlighted in **bold**. The second-best result is underlined.

Dataset	MOO	Margin Rebalancing	Accuracy (%) ↑
FOG Dataset			35.25 ± 1.26
	✓		36.32 ± 2.39
		✓	52.01 ± 2.01
	✓	✓	$\mathbf{56.68 \pm 4.17}$
FBG Dataset			36.20 ± 0.75
	✓		48.78 ± 1.85
		✓	53.66 ± 1.85
	✓	✓	$\mathbf{59.48 \pm 3.03}$
WearGait Dataset			57.38 ± 0.95
	✓		69.95 ± 0.74
		✓	70.86 ± 1.17
	✓	✓	$\mathbf{71.39 \pm 1.09}$

E. Ablation Study

Our approach has two key pieces: a multi-objective optimization (MOO) paradigm and a margin-based class rebalancing strategy. Table V reports an ablation with asynchronous inputs across three datasets. Both components consistently improve accuracy, and the best results arise when they are combined (bold). MOO is especially impactful when class imbalance is milder (FBG and WearGait), delivering gains of 12.58% and 12.57%, respectively, over the non-MOO counterpart. While each component alone outperforms the plain baseline, only their combination achieves the top performance across datasets (with the second-best underlined), highlighting the complementary nature of MOO and margin-based rebalancing.

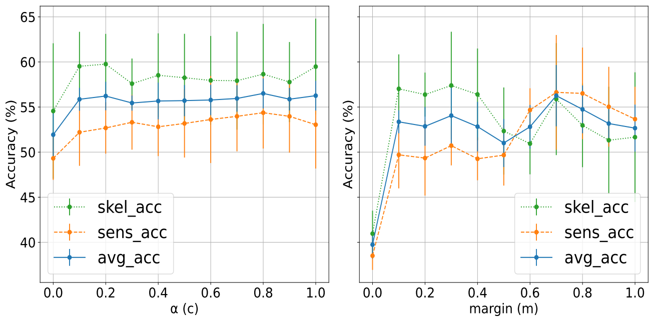

Fig. 5: Performances analysis under asynchronous input settings on the FOG dataset with two hyperparameters: α and m, which respectively represents the MOO coefficient and additive margin coefficient.

F. Hyper-Parameters Analysis

Figure 5 reports the mean accuracy obtained when varying the two free hyper-parameters of our framework: 1. the MOO coefficient α, which dictates how strongly the shared backbone is steered toward the worst-performing modality, and 2. the additive margin m used in our class-rebalancing loss.

a) MOO Coefficient α: The left plot shows a monotonic but saturating trend: raising α from 0 (plain averaging) to ≈ 0.8 yields a gradual $\sim$2–3 pp gain in average accuracy, mainly by lifting the weaker sensor branch while leaving the skeleton branch essentially unchanged. Beyond $\alpha \approx 0.9$ the curve flattens and variance grows, suggesting that very aggressive worst-case weighting offers no further benefit. Hence a moderate range $\alpha \in [0.6, 0.9]$ is sufficient for stable improvements.

b) Additive Margin Coefficient m: The right plot highlights a sharp jump in performance when a small margin is introduced: moving from $m = 0$ to $m \approx 0.3$ boosts all branches, confirming that class-adaptive margins effectively compensate for long-tailed label distributions. Larger margins ($m > 0.5$) lead to oscillations—sensor accuracy continues to rise slightly, whereas skeleton accuracy deteriorates—indicating that excessive separation can impede convergence of the harder (minority) classes.

c) Recommended Setting: Taken together, the analysis shows that the two mechanisms act *orthogonally*: α resolves inter-modality gradient conflict, while m improves intra-class discrimination. Empirically, $\alpha \approx 0.7$–0.9 and $m \approx 0.3$–0.5 strike a good balance between efficacy and stability across datasets.

VI. Conclusion and Discussion

In this paper, we approach PD assessment from the perspective of MOO, aiming to facilitate more practical deployment of AI-assisted solutions in real-world scenarios. Extensive experiments demonstrate that our proposed framework, `TRIP`, not only supports flexible input configurations during both training and inference but also achieves competitive performance. Nevertheless, `TRIP` has certain limitations—for instance, it introduces additional hyper-parameters that may require careful tuning. We plan to further address these challenges in our future work.

References

[1] B. Brakedal, L. Toker, K. Haugarvoll, and C. Tzoulis, "A nationwide study of the incidence, prevalence and mortality of parkinson's disease in the norwegian population," *npj Parkinson's Disease*, vol. 8, no. 1, p. 19, 2022.

[2] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[3] A. Zhao, Y. Liu, X. Yu, X. Xing, and H. Zhou, "Artificial intelligence-enabled detection and assessment of parkinson's disease using multimodal data: A survey," *Information Fusion*, p. 103175, 2025.

[4] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, "Voiceprints analysis using mfcc and svm for detecting patients with parkinson's disease," in *2015 International conference on electrical and information technologies (ICEIT)*. IEEE, 2015, pp. 300–304.

[5] J. Gao, L. Bidulka, M. J. Mckeown, and Z. J. Wang, "Regular rgb-video based eye movement assessment for parkinson's disease," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[6] J. Wang, L. Xue, J. Jiang, F. Liu, P. Wu, J. Lu, H. Zhang, W. Bao, Q. Xu, Z. Ju *et al.*, "Diagnostic performance of artificial intelligence-assisted pet imaging for parkinson's disease: A systematic review and meta-analysis," *NPJ Digital Medicine*, vol. 7, no. 1, p. 17, 2024.

[7] E. Balaji, D. Brindha, and R. Balakrishnan, "Supervised machine learning based gait classification system for early detection and stage classification of parkinson's disease," *Applied Soft Computing*, vol. 94, p. 106494, 2020.

[8] I. El Maachi, G.-A. Bilodeau, and W. Bouachir, "Deep 1d-convnet for accurate parkinson disease detection and severity prediction from gait," *Expert Systems with Applications*, vol. 143, p. 113075, 2020.

[9] J. R. Williamson, B. Telfer, R. Mullany, and K. E. Friedl, "Detecting parkinson's disease from wrist-worn accelerometry in the uk biobank," *Sensors*, vol. 21, no. 6, p. 2047, 2021.

[10] A. Zhao, J. Li, J. Dong, L. Qi, Q. Zhang, N. Li, X. Wang, and H. Zhou, "Multimodal gait recognition for neurodegenerative diseases," *IEEE transactions on cybernetics*, vol. 52, no. 9, pp. 9439–9453, 2021.

[11] R. Gupta, S. Kumari, A. Senapati, R. K. Ambasta, and P. Kumar, "New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in parkinson's disease," *Ageing research reviews*, vol. 90, p. 102013, 2023.

[12] A. P. Creagh, C. Simillion, A. K. Bourke, A. Scotland, F. Lipsmeier, C. Bernasconi, J. van Beek, M. Baker, C. Gossens, M. Lindemann *et al.*, "Smartphone-and smartwatch-based remote characterisation of ambulation in multiple sclerosis during the two-minute walk test," *IEEE journal of biomedical and health informatics*, vol. 25, no. 3, pp. 838–849, 2020.

[13] Y. Pan, Z. Zhou, W. Gong, and Y. Fang, "Sat: A selective adversarial training approach for wifi-based human activity recognition," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 12 706–12 716, 2024.

[14] H. Rao, M. Zeng, X. Zhao, and C. Miao, "A survey of artificial intelligence in gait-based neurodegenerative disease diagnosis," *Neurocomputing*, p. 129533, 2025.

[15] W. Zhang, Z. Yang, H. Li, D. Huang, L. Wang, Y. Wei, L. Zhang, L. Ma, H. Feng, J. Pan *et al.*, "Multimodal data for the detection of freezing of gait in parkinson's disease," *Scientific data*, vol. 9, no. 1, p. 606, 2022.

[16] I. Bavli, A. Ho, R. Mahal, and M. J. McKeown, "Ethical concerns around privacy and data security in ai health monitoring for parkinson's disease: insights from patients, family members, and healthcare professionals," *AI & SOCIETY*, vol. 40, no. 1, pp. 155–165, 2025.

[17] A. Chaudhuri, A. Dutta, T. Bui, and S. Georgescu, "A closer look at multimodal representation collapse," in *Forty-second International Conference on Machine Learning*.

[18] M. Gholami, R. Ward, R. Mahal, M. Mirian, K. Yen, K. W. Park, M. J. McKeown, and Z. J. Wang, "Automatic labeling of parkinson's disease gait videos with weak supervision," *Medical Image Analysis*, vol. 89, p. 102871, 2023.

[19] Z. Zhou, F. Wang, and W. Gong, "i-sample: Augment domain adversarial adaptation models for wifi-based har," *ACM Transactions on Sensor Networks*, vol. 20, no. 2, pp. 1–20, 2024.

[20] J. Feng, Y. He, Y. Pan, Z. Zhou, S. Chen, and W. Gong, "Enhancing fitness evaluation in genetic algorithm-based architecture search for ai-aided financial regulation," *IEEE Transactions on Evolutionary Computation*, vol. 28, no. 3, pp. 623–637, 2024.

[21] C. Fernandes, L. Fonseca, F. Ferreira, M. Gago, L. Costa, N. Sousa, C. Ferreira, J. Gama, W. Erlhagen, and E. Bicho, "Artificial neural networks classification of patients with parkinsonism based on gait," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 2024–2030.

[22] A. Zhao, L. Qi, J. Li, J. Dong, and H. Yu, "A hybrid spatio-temporal model for detection and severity rating of parkinson's disease from gait data," *Neurocomputing*, vol. 315, pp. 1–8, 2018.

[23] B. Vidya and P. Sasikumar, "Gait based parkinson's disease diagnosis and severity rating using multi-class support vector machine," *Applied Soft Computing*, vol. 113, p. 107939, 2021.

[24] H. Zhao, R. Wang, Y. Lei, W.-H. Liao, H. Cao, and J. Cao, "Severity level diagnosis of parkinson's disease by ensemble k-nearest neighbor under imbalanced data," *Expert Systems with Applications*, vol. 189, p. 116113, 2022.

[25] R. A. Abumalloh, M. Nilashi, S. Samad, H. Ahmadi, A. Alghamdi, M. Alrizq, and S. Alyami, "Parkinson's disease diagnosis using deep learning: A bibliometric analysis and literature review," *Ageing research reviews*, p. 102285, 2024.

[26] W. Zeng, C. Yuan, Q. Wang, F. Liu, and Y. Wang, "Classification of gait patterns between patients with parkinson's disease and healthy controls using phase space reconstruction (psr), empirical mode decomposition (emd) and neural networks," *Neural Networks*, vol. 111, pp. 64–76, 2019.

[27] R. Kaur, R. W. Motl, R. Sowers, and M. E. Hernandez, "A vision-based framework for predicting multiple sclerosis and parkinson's disease gait dysfunctions—a deep learning approach," *IEEE journal of biomedical and health informatics*, vol. 27, no. 1, pp. 190–201, 2022.

[28] B. Vidya and P. Sasikumar, "Parkinson's disease diagnosis and stage prediction based on gait signal analysis using emd and cnn–lstm network," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105099, 2022.

[29] C. Zhong and W. W. Ng, "A robust frequency-domain-based graph adaptive network for parkinson's disease detection from gait data," *IEEE Transactions on Multimedia*, vol. 25, pp. 7076–7088, 2022.

[30] Y. Xia, Z. Yao, Q. Ye, and N. Cheng, "A dual-modal attention-enhanced deep learning network for quantification of parkinson's disease characteristics," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 1, pp. 42–51, 2019.

[31] V. Skaramagkas, A. Pentari, Z. Kefalopoulou, and M. Tsiknakis, "Multi-modal deep learning diagnosis of parkinson's disease—a systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2399–2423, 2023.

[32] V. Radu, C. Tong, S. Bhattacharya, N. D. Lane, C. Mascolo, M. K. Marina, and F. Kawsar, "Multimodal deep learning for activity and context recognition," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 1, no. 4, pp. 1–27, 2018.

[33] P. Kumar, S. Mukherjee, R. Saini, P. Kaushik, P. P. Roy, and D. P. Dogra, "Multimodal gait recognition with inertial sensor data and video using evolutionary algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 5, pp. 956–965, 2018.

[34] F. M. Castro, M. J. Marin-Jimenez, N. Guil, and N. Pérez de la Blanca, "Multimodal feature fusion for cnn-based gait recognition: an empirical comparison," *Neural Computing and Applications*, vol. 32, pp. 14 173–14 193, 2020.

[35] K. Hu, Z. Wang, K. A. E. Martens, M. Hagenbuchner, M. Bennamoun, A. C. Tsoi, and S. J. Lewis, "Graph fusion network-based multimodal learning for freezing of gait detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1588–1600, 2021.

[36] Y. Cui and Y. Kang, "Multi-modal gait recognition via effective spatial-temporal feature fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 949–17 957.

[37] A. Navon, A. Shamsian, G. Chechik, and E. Fetaya, "Learning the pareto front with hypernetworks," *arXiv preprint arXiv:2010.04104*, 2020.

[38] M. Ruchte and J. Grabocka, "Scalable pareto front approximation for deep multi-objective learning," in *2021 IEEE international conference on data mining (ICDM)*. IEEE, 2021, pp. 1306–1311.

[39] N. Dimitriadis, P. Frossard, and F. Fleuret, "Pareto manifold learning: Tackling multiple tasks via ensembles of single-task models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 8015–8052.

[40] ——, "Pareto low-rank adapters: Efficient multi-task learning with preferences," *arXiv preprint arXiv:2407.08056*, 2024.

[41] W. Chen and J. T. Kwok, "Efficient pareto manifold learning with low-rank structure," *arXiv preprint arXiv:2407.20734*, 2024.

[42] A. Tang, L. Shen, Y. Luo, S. Liu, H. Hu, and B. Du, "Towards efficient pareto set approximation via mixture of experts based model fusion," *arXiv preprint arXiv:2406.09770*, 2024.

[43] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[44] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5824–5836, 2020.

[45] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, "Conflict-averse gradient descent for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 878–18 890, 2021.

[46] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi, G. Chechik, and E. Fetaya, "Multi-task learning as a bargaining game," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 428–16 446.

[47] H. Ban and K. Ji, "Fair resource allocation in multi-task learning," *arXiv preprint arXiv:2402.15638*, 2024.

[48] F. Heidarivincheh, R. McConville, C. Morgan, R. McNaney, A. Masullo, M. Mirmehdi, A. L. Whone, and I. Craddock, "Multimodal classification of parkinson's disease in home environments with resiliency to missing modalities," *Sensors*, vol. 21, no. 12, p. 4133, 2021.

[49] S. Liu, T. Kimura, D. Liu, R. Wang, J. Li, S. Diggavi, M. Srivastava, and T. Abdelzaher, "Focal: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47 309–47 338, 2023.

[50] Z. Lv, T. Pan, C. Si, Z. Chen, W. Zuo, Z. Liu, and K.-Y. K. Wong, "Rethinking cross-modal interaction in multimodal diffusion transformers," *arXiv preprint arXiv:2506.07986*, 2025.

[51] S. Mo and P. Morgado, "Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 186–27 196.

[52] C. Ribeiro De Souza, R. Miao, J. Ávila De Oliveira, A. Cristina De Lima-Pardini, D. Fragoso De Campos, C. Silva-Batista, L. Teixeira, S. Shokur, B. Mohamed, and D. B. Coelho, "A public data set of videos, inertial measurement unit, and clinical scales of freezing of gait in individuals with parkinson's disease during a turning-in-place task," *Frontiers in Neuroscience*, vol. 16, p. 832463, 2022.

[53] T. K. F. Shida, T. M. Costa, C. E. N. de Oliveira, R. de Castro Treza, S. M. Hondo, E. Los Angeles, C. Bernardo, L. dos Santos de Oliveira, M. de Jesus Carvalho, and D. B. Coelho, "A public data set of walking full-body kinematics and kinetics in individuals with parkinson's disease," *Frontiers in Neuroscience*, vol. 17, p. 992585, 2023.

[54] K. Kontson *et al.*, "Wearables for gait in parkinson's disease and age-matched controls," *Synapse by SAGE Bionetworks*, 2024.

[55] K.-K. Maninis, I. Radosavovic, and I. Kokkinos, "Attentive single-tasking of multiple tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1851–1860.

[56] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.