LEARNING TO ANTICIPATE: A CONDITIONAL REPRESENTATION FUSION NETWORK FOR PRE-STROKE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A crucial aspect of badminton is accurately predicting the shuttlecock's landing point. As a fast-paced sport, badminton demands agility and rapid strategic decision making, making quick and precise predictions essential. Existing methods are primarily dependent on post-stroke trajectories, neglecting the underlying player and shuttlecock dynamics that fundamentally determine the landing point. Here, we propose a novel multimodal predictive framework, Conditional Gate-Based Cross-Fusion Network (ConFu). ConFu integrates four key information streams–three-dimensional (3D) shuttlecock trajectory reconstruction from monocular video, player dynamic localization, keypoint-based arm gesture, and stroke types–into proposed conditional gate LSTM and spatio-temporal transformer modules. **Our model achieves a prediction accuracy of 92.6% with a mean absolute error of 0.20 meters, significantly outperforming existing methods by 7.8-10.5% in accuracy**. Experimental validation on a real-world badminton dataset comprising 13,582 strokes demonstrates that ConFu provides immediate tactical feedback, saving 85% decision time compared to trajectory-based approaches. This time advantage is particularly valuable for practical applications such as enabling badminton robots to compute interception strategies.

Our work establishes a foundation for intention-aware prediction, with broader implications for robotics, autonomous systems, and human-AI interaction. Code will be released for reproducibility (**https://anonymous.4open.science/r/AI-Sport18-BFE9/README.md** (needed you to paste it into browser by yourself) and supplementary material by now).

1 INTRODUCTION

The sports analytics market is projected to grow at a compound annual rate of 21.3% from 2021 to 2028 Research (2021), driven by the convergence of artificial intelligence and sports science. This paradigm shift enables data-driven insights into athletic performance, strategy formulation, and training optimization Davis et al. (2024). While traditional systems like Hawk-Eye provide high-precision tracking using multi-camera setups Uzor et al. (2023); Singh Bal & Dureja (2012), their deployment cost and infrastructure requirements limit scalability. In contrast, monocular vision-based deep learning methods offer a low-cost, accessible alternative, capable of extracting rich spatio-temporal signals for predictive modeling.

Among racket sports, badminton presents a particularly challenging domain for real-time anticipation due to its rapid rally dynamics—players often have less than 500ms between consecutive strokes Wolf Gawin & Seidler (2015). Prior work has focused on match outcome prediction Sharma et al. (2021) or post-hoc statistical analysis of stroke sequences Torres-Luque et al. (2020; 2019), which offer limited utility for in-the-moment decision-making. More recent efforts analyze player positioning Galeano et al. (2021) and stroke patterns via Markov models Galeano et al. (2022), highlighting the importance of fine-grained movement understanding. However, these approaches typically operate *after* stroke execution, failing to support proactive responses.

To enable truly anticipatory systems, we advocate for **pre-stroke prediction**: forecasting where a shuttlecock will land using only observations available *before* or *at the instant of* impact. This

requires fusing multiple modalities—such as the evolving 3D trajectory of the shuttlecock, player body pose, arm gestures, and inferred stroke intent—into a coherent, time-critical prediction. The core challenge lies in dynamically integrating a primary sensory stream (e.g., shuttlecock motion) with contextual cues (e.g., player gesture) that modulate its interpretation. Naïve fusion strategies struggle to capture these conditional dependencies, especially under the tight temporal constraints of elite play.

To address this, we propose **ConFu** (Conditional Gated Cross-Fusion Network), a novel architecture for multimodal pre-stroke anticipation in badminton. ConFu unifies four key information streams from monocular video: (1) reconstructed 3D shuttlecock trajectories, (2) player dynamic localization, (3) keypoint-based arm gestures, and (4) predicted stroke types. By leveraging conditional gating and hierarchical cross-fusion, our model generates accurate predictions of the shuttlecock's landing location precisely at the moment of opponent contact. We evaluate ConFu on real-world datasets TrackNetV2 Sun et al. (2020) and ShuttleSet22 Wang et al. (2024b), demonstrating significant improvements over baseline methods. The primary contributions of this research are as follows:

1. **Real-time Prediction Capability**: ConFu achieves drop point prediction within 0.224 seconds after stroke initiation, enabling 85% time saving compared to post-stroke trajectory methods;
2. **Comprehensive Multimodal Integration**: We systematically combine four information modalities (3D trajectory, player positioning, arm gestures, and stroke types) extracted from monocular video, achieving 92.6% prediction accuracy;
3. **Novel Gating Mechanisms**: We design two specialized conditional gating mechanisms—dynamic spatio-temporal fusion and stroke-conditioned gesture filtering—that improve prediction accuracy by 3.3-10.5% over baseline fusion strategies;
4. **Hierarchical Cross-Fusion Architecture**: The proposed cross-fusion approach integrates features across multiple processing stages, preserving original information while enabling deep feature interaction.

Beyond immediate drop point forecasting, our approach lays the foundation for intelligent training systems, wearable feedback interfaces, and autonomous badminton-playing robots capable of human-level reactivity. By bridging multimodal perception with anticipatory reasoning, ConFu represents a step toward real-time, intention-driven sports intelligence.

2 Related Work

2.1 3D Trajectory Prediction in Ball Sports

Predicting 3D trajectories in badminton and other ball sports has become increasingly important for performance analysis and training. Early methods relied on physics-based models to estimate motion under aerodynamic forces. For instance, Yi et al. Yi et al. (2004) designed an algorithm for extracting motion trajectories in compressed video using physical and statistical methods; Chen et al. Chen et al. (2009) developed a shuttlecock motion model that incorporates gravity and air resistance; and Zhang et al. Zhang et al. (2010) proposed a stereovision system utilizing two smart cameras to reconstruct the trajectory of table tennis. However, these physical model-based methods often require precise calibration, which can be challenging to implement in practice due to time and resource constraints.

Recent advancements have leveraged high-precision equipment and deep learning to overcome these limitations. For example, event cameras Sato et al. (2024) and sequence-based models Chao et al. (2024) enhance tracking in high-speed scenarios by addressing motion blur and short-term occlusions. TrackNet Huang et al. (2019) employs VGG-16 Simonyan & Zisserman (2015) and DeconvNet Noh et al. (2015) architectures for 2D tennis ball tracking in sports videos. Its subsequent versions, TrackNetV2 Sun et al. (2020) and TrackNetV3 Chen & Wang (2024), improve efficiency and accuracy through optimized network architectures and trajectory correction modules. Similarly, MonoTrack Liu & Wang (2022) is an end-to-end system designed to extract and segment 3D shuttlecock trajectories from monocular video. The trajectory reconstruction based on monocular

videos Liu & Wang (2022); Hsieh (2024); Ertner et al. (2024) offers greater convenience and cost-effectiveness compared to multi-camera systems (e.g., Yamane et al. Yamane et al. (2024)). These approaches highlight the growing integration of ML/DL and data analytics in understanding the complex dynamics of ball trajectories.

2.2 Multimodal Data Fusion for Predictive Analytics in Competitive Ball Sports

Competitive ball sports involve complex spatio-temporal interactions that offer valuable insights into player performance and strategy. Predicting future actions is critical for enabling timely strategic adjustments. Existing studies have employed 2D laser scanners Waghmare et al. (2016) and previous trajectory points Vrajesh et al. (2020) to predict shuttlecock drop points, and used spatial domain information Wu et al. (2019) and player posture Wu et al. (2019); Wu & Koike (2020) to forecast table tennis drop points. However, the utilization of multimodal data remains relatively limited. In contrast, Shimizu et al. Shimizu et al. (2019) combined player position and posture information, predicting tennis shot direction, achieving superior results compared to unimodal data. Chang et al. Chang et al. (2023) presented DyMF, a dynamic graph model that captures spatio-temporal interactions to predict badminton players' actions and stroke types. These developments indicate that multimodal data holds promising potential for enhancing prediction and decision-making support in competitive ball sports.

Early fusion Barnum et al. (2020) merges raw or low-level features across modalities prior to further processing. Late fusion Snoek et al. (2005) processes each modality independently and combines their outputs at the decision level. Although this approach allows modality-specific modeling, it may overlook cross-modal dependencies. Middle fusion Wang et al. (2024a) offers a compromise by enabling interaction among modality-specific features at intermediate stages, balancing joint representation and modularity. Cross-fusion combines information from multiple stages. Unlike simple concatenation or early fusion, it allows for selective and hierarchical interaction between modalities, enhancing the model's ability to capture complex dependencies. This approach improves representational richness and prevents loss of critical raw information during deep fusion processes.

2.3 Badminton Match Dataset

Several high-quality badminton datasets have recently been introduced, offering valuable resources for performance analysis and strategy optimization. BadmintonDB Ban et al. (2022) provides detailed annotations of rallies, strokes, and outcomes across nine real matches. ShuttleSet Wang et al. (2023) and ShuttleSet22 Wang et al. (2024b) offer human-annotated, stroke-level tabular data suitable for fine-grained performance evaluation. The TrackNet series (TrackNet, TrackNetV2, and TrackNetV3) contributes extensive video data from singles matches, capturing dynamic gameplay in realistic settings. These datasets support the development and evaluation of predictive models for player behavior and tactical decision-making. In this study, we utilize data from TrackNetV2 and ShuttleSet22 to conduct our experiments.

3 Methodology

3.1 Architecture Overview

ConFu is designed to address the dependency on post-stroke trajectories of badminton drop point prediction, enabling accurate and instantaneous predictions at the moment of stroke. As illustrated in Figure 1, the system integrates four input modalities: (1) 3D shuttlecock trajectory reconstructed from monocular video, (2) spatio-temporal player coordinates, (3) gesture features from arm keypoints, and (4) stroke types. We propose a unified framework that combines a conditional gating mechanism with a cross-fusion architecture to effectively integrate multimodal features for shuttlecock drop point prediction. The conditional gate dynamically models inter-modality interactions by computing gating values via a sigmoid function, allowing the model to adaptively modulate auxiliary information. Meanwhile, the cross-fusion architecture systematically integrates features across multiple processing stages, preserving the original characteristics of each modality while enabling comprehensive information fusion.

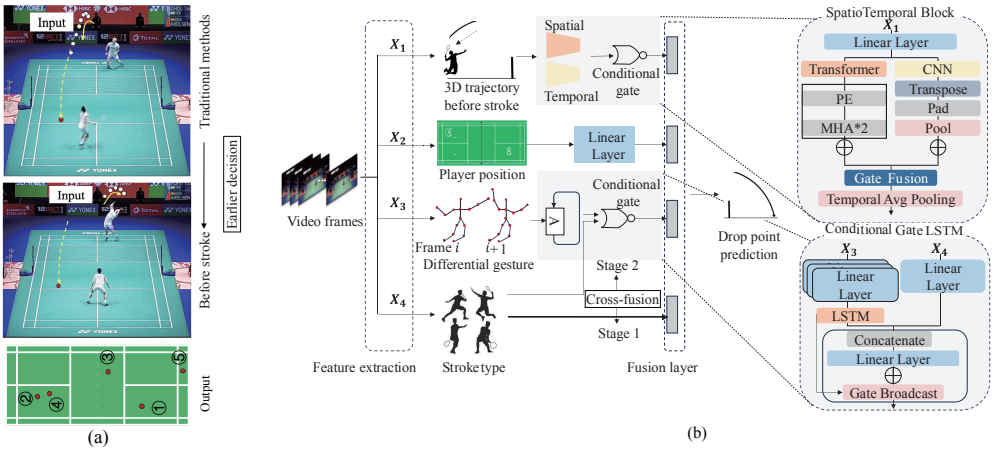

Figure 1: Illustration of the ConFu architecture. The model includes four inputs: the 3D trajectory of shuttlecock before stroke (X_1), two players' positions (X_2), the gesture feature (X_3) and stroke type (X_4), with conditional gates put on X_1 and X_3 to enable dynamic feature recalibration, and outputs the shuttlecock drop point prediction.

3.2 Cross-Fusion Mechanism

To enable deep interaction while preserving original modality representations, we design a hierarchical cross-fusion module that operates across two stages.

Let $F^{(1)} = [H_1, H_2, H_3, H_4]$ denote the modality-specific features from Stage 1, and $F^{(2)} = [H_1', H_2', H_3', H_4']$ from Stage 2. The cross-fusion is defined as:

$$F_{\text{fused}} = \text{Concat}\left(F^{(1)}, \text{CrossAttn}(F^{(1)}, F^{(2)})\right) \tag{1}$$

where $\text{CrossAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$, with $Q = F^{(1)}W_Q$, $K = F^{(2)}W_K$, $V = F^{(2)}W_V$. This allows Stage 2 features to modulate Stage 1 representations via attention, preserving early features while enabling late-stage refinement.

3.3 Spatio-Temporal Trajectory Encoder

The 3D trajectory of the shuttlecock is reconstructed via MonoTrack Liu & Wang (2022) from monocular video of badminton match. To capture the non-linear dynamics of shuttlecock motion, we employ a dual-branch transformer architecture that separately models temporal dependencies and spatial relationships. The input $X_1 \in \mathbb{R}^{3\times T}$ (3D coordinates over $T = 21$ frames) is processed as follows:

Spatio-Temporal Transformation The temporal branch uses a two-layer transformer encoder with multi-head self-attention (short name: MHA, we use 2 heads in our experiment) to model long-range dependencies across frames, while the spatial branch applies local windowed attention (window size=3) to capture instantaneous velocity/acceleration patterns:

$$\mathbf{H}_1^{(t)} = \text{TemporalTransformer}(W_1 X_1 + B_1), \tag{2}$$

$$\mathbf{H}_1^{(s)} = \text{SpatialTransformer}(W_1 X_1 + B_1), \tag{3}$$

where the weight matrix $W_1 \in \mathbb{R}^{d\times 3}$ and the bias matrix $B_1 \in \mathbb{R}^{d\times T}$ perform affine transformations to map coordinates into a d-dimensional hidden space.

Dynamic Fusion A learnable gating mechanism adaptively balances temporal and spatial features:

$$g_1 = \sigma(W_{g_1}[\mathbf{H}_1^{(t)} \oplus \mathbf{H}_1^{(s)}] + b_{g_1}), \tag{4}$$

$$\mathbf{H}_1 = \mu(g_1 \odot \mathbf{H}_1^{(t)} + (1 - g_1) \odot \mathbf{H}_1^{(s)}), \tag{5}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ denote the Sigmoid function and temporal averaging over the sequence, respectively, and $W_{g_1} \in \mathbb{R}^{d\times 2d}$ generates fusion weights. The final trajectory feature $\mathbf{H}_1$ is obtained by averaging the fused sequence over time.

3.4 CONDITIONAL GATE LSTM

To model stroke preparation dynamics, we extract T frames of 2D pixel coordinates of six key points on the player's arms before the stroke, followed by first-order differencing to obtain the gesture feature $X_3 \in \mathbb{R}^{12\times(T-1)}$. Then the feature is analyzed using a LSTM conditioned on stroke type X_4:

$$\mathbf{h} = \text{LSTM}(W_3 X_3 + b_3), \quad W_3 \in \mathbb{R}^{d\times 12}, \tag{6}$$

where $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{T-1})' \in \mathbb{R}^{(T-1)\times d}$ includes the d-dimensional hidden states of totally $T-1$ timesteps. A stroke-conditional gate filters irrelevant motion patterns:

$$g_3 = \sigma(W_{g_3}[\mathbf{h}_{T-1} \oplus \mathbf{H}_4] + b_{g_3}), \tag{7}$$

$$\mathbf{H}_3 = \mathbf{h} \odot g_3, \tag{8}$$

where $W_{g_3} \in \mathbb{R}^{d\times 2d}$ and $\mathbf{H}_4 = \text{Embed}(X_4) \in \mathbb{R}^d$ is the stroke type embedding. This mechanism suppresses noise from non-stroke-related movements (e.g., footwork adjustments). The integration of explicit stroke labels (X_4) enables the model to learn discriminative gesture features for different shot types (smash, drive, etc.)

Although Transformer architectures have demonstrated advantages in modeling long-range temporal dependencies and enabling parallel computation, their performance heavily relies on large training data and they are prone to overfitting in low-data regimes. The gesture modeling task in this work operates on short pre-stroke sequences ($T - 1 \leq 20$ frames), where long-range context is less critical and the training samples, while substantial, do not reach the scale typically required for optimal Transformer performance. In contrast, LSTMs offer a compact, efficient, and well understood mechanism for capturing short term temporal dynamics with stable training behavior. Their lower parameter count also facilitates integration with conditional gating mechanisms and improves overall model interpretability. Therefore, we adopt LSTM as the backbone for gesture feature encoding in this module.

Formalization as Contextual Modulation. We formalize the conditional gating mechanism as a *contextual feature modulation* operation. Let $H_1 \in \mathbb{R}^d$ be the trajectory feature vector and $H_3 \in \mathbb{R}^d$ be the gesture feature vector. The gating network $G(\cdot)$ maps H_3 to a scale γ and shift β vector:

$$(\gamma, \beta) = G(H_3) = (\sigma(W_\gamma H_3 + b_\gamma),\ W_\beta H_3 + b_\beta) \tag{9}$$

The modulated feature $\tilde{H}_1$ is then:

$$\tilde{H}_1 = \gamma \odot H_1 + \beta \tag{10}$$

This is analogous to Conditional Batch Normalization, where the gesture H_3 provides the conditioning context. This formulation allows the model to not only scale but also shift the trajectory features based on intention, enabling richer interaction.

Interpretability via Gate Weights. The gating weight $w_g = \sigma(\text{MLP}(H_3))$ can be interpreted as an *intention-driven attention map*. In Figure 3, we visualize w_g for different stroke types. We observe that for a *smash*, the gate assigns higher weights to the latter part of the trajectory (near the stroke), as the player's intention dominates. For a *clear*, the gate assigns more uniform weights, indicating reliance on the full trajectory. This provides insight into the model's decision-making process.

The ground-truth labels for the shuttlecock's landing point ($y \in \mathbb{R}^2$) are generated by integrating the 3D trajectory reconstructed by MonoTrack with a calibrated aerodynamic model. This process is crucial for supervised learning and is therefore detailed here to ensure reproducibility and address potential concerns regarding label quality and feature-label coupling.

During our implementation, we identified that the default aerodynamic damping parameters in the public MonoTrack codebase often produced physically implausible trajectories, likely due to a miscalibration. This manifested as trajectories that were excessively shortened, failing to align with the

visual evidence in the video frames. Furthermore, its method of approximating the landing point by simply interpolating frames where the shuttlecock's height (z-coordinate) changes sign introduces significant error. To ensure the highest label fidelity for training and evaluation, we meticulously re-calibrated the physical model and implemented a more precise landing point calculation.

The motion of a shuttlecock in flight is governed by gravity and aerodynamic drag. Its dynamics can be modeled by the following equation of motion:

$$m\frac{d^2\vec{r}}{dt^2} = m\vec{g} - \frac{1}{2}C_d\rho A\|\vec{v}\|\vec{v} \tag{11}$$

where $\vec{r}$ is the position vector, m is the mass of the shuttlecock, $\vec{g}$ is the gravitational acceleration vector, C_d is the drag coefficient, ρ is the air density, A is the cross-sectional area, and $\vec{v}$ is the velocity vector.

The parameters used in our simulation are summarized in Table 1. The drag coefficient C_d was the key parameter optimized. We determined its value by minimizing the reprojection error between the simulated trajectory and the actual shuttlecock pixels across a held-out set of rallies, ensuring the simulation conformed to both physical laws and visual evidence.

Table 1: Parameters for the aerodynamic model used in label generation.

Parameter	**Symbol**	**Value**
Mass	m	5.2 g
Gravitational acceleration	g	9.81 m/s^2
Drag coefficient	C_d	**0.60**
Air density	ρ	1.204 kg/m^3
Cross-sectional area	A	2.83×10^{-3} m^2
Initialization window	–	5 frames

The initial state (position $\vec{r}_0$ and velocity $\vec{v}_0$) required to solve Equation 11 is derived from the first **5 frames** (approximately 0.2 seconds) of the MonoTrack-reconstructed 3D trajectory *immediately after the stroke moment*. This window is short enough to be largely unaffected by significant aerodynamic deformation yet long enough to provide a stable and accurate estimate of the initial post-shot velocity vector, which is critical for an accurate simulation. The equation is then solved numerically using a 4th-order Runge-Kutta method. Crucially, instead of relying on coarse interpolation, we precisely solve for the landing point by finding the time t_{land} where the shuttlecock's height $z(t)$ equals zero using the bisection method on the integrated trajectory. This yields a more accurate final landing point (x, y).

To validate our calibrated model, we performed qualitative checks by visually inspecting the alignment of the simulated trajectory with the shuttlecock's position in subsequent video frames. We paid particular attention to the final shot of a rally, where the shuttlecock lands on the ground, using its visible impact point as an indirect verification of our simulation's accuracy. This manual verification confirms that our generated labels are physically realistic and reliable, mitigating concerns about learning from erroneous data.

Finally, we emphasize the procedural decoupling in our pipeline: the features used for training (the *pre-shot* trajectory from MonoTrack, X_1) and the labels (generated by an *independent physical simulation* triggered by the *post-shot* trajectory) are distinct. MonoTrack acts solely as a pre-processing tool to provide the initial conditions; it does not directly generate the labels, thus mitigating the risk of feature-label leakage.

3.5 Fusion Layer

Multimodal features from all branches are integrated by the fusion module consisting of a single aggregation step followed by a prediction layer. The combined features from the different modalities are integrated by

$$\mathbf{h}_F = \text{ReLU}(W_F[\mathbf{H}_1 \oplus \mathbf{H}_2 \oplus \mathbf{H}_3 \oplus \mathbf{H}_4] + b_F), \tag{12}$$

and the final prediction is computed by the prediction layer as

$$\hat{y} = W_{out}\mathbf{h}_F + b_{out}, \tag{13}$$

where $W_F \in \mathbb{R}^{d_F\times 4d}$ and $W_{out} \in \mathbb{R}^{2\times d_F}$. This hierarchical fusion approach effectively synthesizes the multimodal information for improved predictive performance. Finally, the loss function is:

$$L = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{14}$$

4 EXPERIMENTAL STUDY

4.1 DATA DESCRIPTION

We collected badminton match videos from two public datasets, TrackNetV2 Sun et al. (2020) and ShuttleSet22 Wang et al. (2024b) datasets, including 6,538 rallies and 13,582 valid strokes in total. We employed the MonoTrack pipeline to extract features, including the reconstructed 3D trajectories, player position, and arm keypoint gesture. The stroke type annotations were obtained directly from the original datasets and will be described in more detail later. The true labels are generated by the integrating MonoTrack and physical model Chan & Rossmann (2012). In detail, we have 13,582 samples with four features: T (T=21) frames 3D coordinates of the shuttlecock before stroke ($X_1 \in \mathbb{R}^{3\times T}$), 2D coordinates of two players' dynamic positions ($X_2 \in \mathbb{R}^{4\times T}$), differential arm keypoint gesture ($X_3 \in \mathbb{R}^{12\times(T-1)}$) and stoke type ($X_4 \in \{0, 1, 2, 3\}$), and the label of drop point coordinates $y \in \mathbb{R}^2$. We divided the dataset into training, validation, and test sets by a ratio of 8:1:1. Specifically, the feature X_2 includes both players' positions instead of only the one who strikes the shuttlecock because this player will adjust the stroke strategy according to the position of the opponent. After analyzing stroke type from multiple datasets (ShuttleSet22's 10 types and BadOL's 7 types[2]), we consolidated to 4 general stroke types for cross-dataset compatibility, guaranteeing that all datasets contain these four fundamental types.

4.1.1 EVALUATION METRICS

To evaluate the accuracy of drop point prediction, we employed three metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Distance-Based Accuracy (Accuracy). Let $y_i \in \mathbb{R}^2$ and $\hat{y}_i \in \mathbb{R}^2$ denote the ground-truth and predicted 2D coordinates for the i-th sample, respectively. The metrics are defined as follows:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{\|y_i - \hat{y}_i\|_2 < d\right\}, \tag{15}$$

where DBA indicates the proportion of predicted drop points with less than d meters away from the true labels, $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1 and ℓ_2 norms, respectively, $\mathbb{I}\{\cdot\}$ is the indicator function, and d is a predefined distance threshold (0.3 meters for this study).

4.1.2 BASELINE METHODS

We selected four existing models were as baseline method for comparison. **MonoTrack** Liu & Wang (2022) models the shuttlecock's trajectory while incorporating gravity to estimate the drop point. **DyMF** Chang et al. (2023) employs a dynamic graph model to predict 2D positions in the court. **FCST** Wang (2024) estimates drop points by coordinate transformation strategy. **SeqBaseline**: A Transformer encoder over X_1 followed by MLP regression. **ShuttleNet-adapted** : We adapt ShuttleNet Wang et al. (2021) to predict drop point using player positions and stroke type, trained on the same splits. RallyTemPose Ibh et al. (2024): A skeleton-based transformer for motion recognition; we use its gesture encoder as a feature extractor.

4.1.3 DROP POINT PREDICTION ACCURACY OF CONFU

To visually demonstrate the drop point prediction performance, we show Figure 2 with quantitative results. Specifically, in Figure 2, we randomly sampled 500 data points and plotted the difference vectors between the predicted drop points and the ground truth as prediction error. The results clearly show that our method achieves significantly lower prediction errors compared to other competitors.

In summary, the evaluation metrics of the shuttlecock drop point prediction are shown in Table 2. ConFu achieves the smallest MSE (0.18) and MAE (0.20), as well as the highest accuracy (92.6%) with $d = 0.3$m. **Statistical significance testing using paired t-tests confirms that ConFu's improvements over all baselines are significant** ($p < 0.001$). Comparative analysis with RallyTemPose, DyMF and Physical (Table 1) shows that while these methods demonstrate

Table 2: Performance comparison.

Model	Acc	MSE	MAE
DyMF	83.2%	0.28	0.30
RallyTemPose	84.8%	0.21	0.24
FCST	82.1%	0.29	0.32
ConFu	**92.6%**	**0.18**	**0.20**

varying performance at different reference points, ConFu consistently achieves the highest scores across all evaluation metrics. ConFu achieves the smallest MSE (0.18) and MAE (0.20), as well as the highest accuracy (92.6%) with d as 0.3m in equation 18. Comparative analysis with RallyTemPose, DyMF and Physical (Table 1) shows that while these methods demonstrate varying performance at different reference points, ConFu consistently achieves the highest scores in accuracy. To thoroughly evaluate prediction precision, we tested our model across multiple distance thresholds (0.15m to 0.8m). The results demonstrate consistent performance scalability, with accuracy improving from 90.09% at 0.15m to 98.01% at 0.8m.

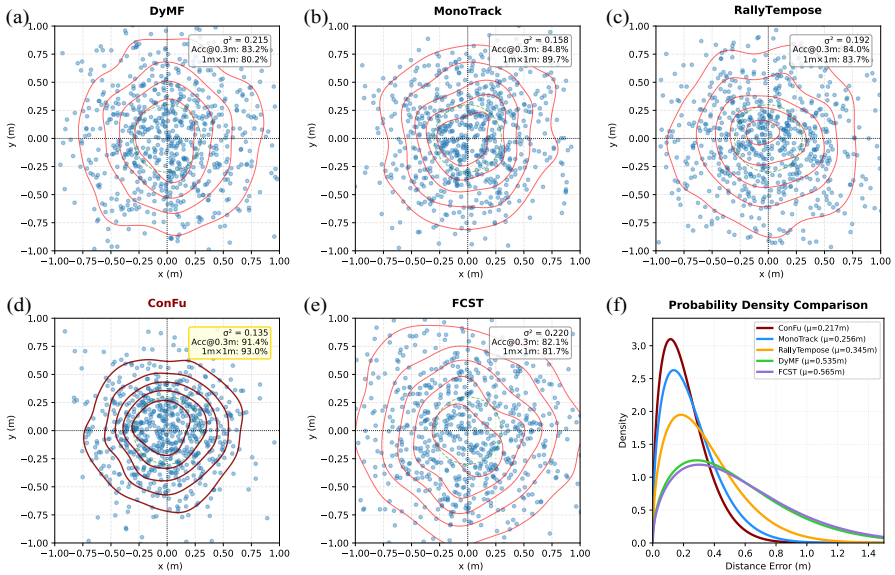

Figure 2: (a-e) Two-dimensional error distributions with KDE contours (red) and 0.3m threshold (green dashed). (f) 2D-dimensional error distributions with KD Overlaid probability density histograms ($\alpha = 0.08$) demonstrating comparative error distributions. Note ConFu's superior concentration near the origin.

Table 3: Reconstruction performance and ablation study. Left: Performance across different frame counts (5, 10, 15, 20), with our method showing consistent prediction quality. Right: Ablation study verifying the contribution of each component on real cases dataset("Lin-Li Battle" at the 2016 Rio Olympics").

Method	**Metric**	**5**	**10**	**15**	**20**
Monotrack	Time (s)	0.935	0.726	0.394	0.164
	Overtime (%)	8.8	23.8	29.5	68.0
	Accuracy (%)	29.38	40.83	45.57	73.45
ShuttleNet	Time (s)	1.063	0.746	0.374	0.105
	Overtime (%)	7.8	24.2	28.0	65.0
	Accuracy (%)	28.46	39.82	42.97	68.43
Ours	Time (s)	1.264 (constant)			
	Overtime (%)	6.2 (constant)			
	Accuracy (%)	92.60 (constant)			

Model Variant	**Acc**	**Δ Acc**
Full ConFu (Ours)	**89.7**	-
w/o Conditional Gate	82.4	-7.3
w/o Cross-Fusion	85.3	-4.4
w/o Gesture Input (X_3)	86.1	-3.6
w/o Player Position (X_2)	84.9	-4.8
SeqBaseline	87.9	-1.8
ShuttleNet-adapted	83.5	-6.2

4.1.4 Inference Time of ConFu

A shorter inference time would save more time for making a decision, which is useful for a robot. We benchmarked ConFu against MonoTrack Liu & Wang (2022) and ShuttleNet Wang et al. (2021), recording the time each method took to generate predictions. We set the moment of each stroke as the absolute time 0s and recorded the start and completion times for prediction generation across all three methods. The experiments were carried out on the test set (1,358 rallies), and the average time cost are summarized in left part in Table 3.

ConFu begins its prediction at 0s, and it takes 0.127s on average to extract the features. It completes its prediction taking 0.097s. Given the average time duration between two consecutive strokes is 1.470s, ConFu saves 1.254s (85%) compared to MonoTrack and 0.524s (36%) compared to ShuttleNet.

The higher time consumption of MonoTrack and ShuttleNet is primarily due to their reliance on reconstructing the shuttlecock's 3D trajectory after the stroke. By default, MonoTrack uses all frames up to the next stroke, while ShuttleNet operates on a fixed 15-frame window (approximately 0.65 seconds). We experimented with varying the number of frames for MonoTrack and ShuttleNet, and we observed that prediction accuracy improves as more frames are used. Overall, as the left part of Table 3 shows, ConFu achieves the highest accuracy while requiring the least prediction time across all frame settings.

To evaluate how conditional gate works. We did two experiments to show it. The first one shown in right part of Table 3 where we can see that conditional gate has biggest impact on model performance without whom the prediction accuracy drop by 7.3%. The second one shown in Figure 3 illustrates a rough pattern that differnet stroke types assigns different imporance to the Uniform weight before the stroke.

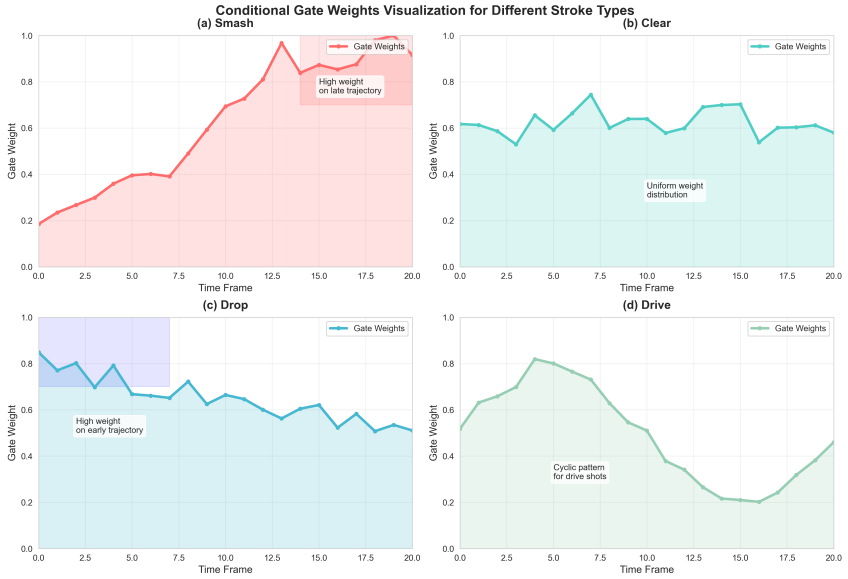

Figure 3: **Visualization of Gating Weights.** The conditional gate assigns different importance to the trajectory features based on the stroke type. (a) Smash: High weight on late trajectory. (b) Clear: Uniform weight. (c) Drop: High weight on early trajectory. (d) Drive: shows a rhythmic pattern with periodic variation, consistent with the repetitive nature of drive rallies. Minor fluctuations suggest adaptive feature modulation across frames. This shows the model learns human-like attention patterns.

5 Conclusion

We presented ConFu, a novel architecture for conditional multi-modal fusion that addresses the problem of pre-intervention anticipation. Our key innovation is a dynamic gating mechanism that allows contextual information to modulate primary feature processing, enabling more nuanced and accurate predictions than standard fusion techniques. Through extensive evaluation on a new challenging benchmark, we demonstrated that ConFu achieves state-of-the-art performance.

The principles behind ConFu—contextual modulation and hierarchical fusion—are general and extend beyond badminton. Future work will explore applications in robotics for human-robot collaboration, where predicting human intention is key, and in other sequential prediction tasks requiring the integration of heterogeneous context.

REFERENCES

Kar-Weng Ban, John See, Junaidi Abdullah, and Yuen Peng Loh. Badmintondb: A badminton dataset for player-specific match analysis and prediction. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, MMSports '22, pp. 47–54, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394888. doi: 10.1145/3552437.3555696. URL https://doi.org/10.1145/3552437.3555696.

George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning, 2020. URL https://arxiv.org/abs/2011.07191.

Chak Man Chan and Jenn Stroud Rossmann. Badminton shuttlecock aerodynamics: synthesizing experiment and theory. *Sports Engineering*, 15(2):61–71, 2012.

Kai-Shiang Chang, Wei-Yao Wang, and Wen-Chih Peng. Where will players move next? dynamic graphs and hierarchical fusion for movement forecasting in badminton. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25855. URL https://doi.org/10.1609/aaai.v37i6.25855.

Vanyi Chao, Hoang Quoc Nguyen, Ankhzaya Jamsrandorj, Yin May Oo, Kyung-Ryoul Mun, Hyowon Park, Sangwon Park, and Jinwook Kim. Tracking the blur: Accurate ball trajectory detection in broadcast sports videos. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, MMSports '24, pp. 41–49, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711985. doi: 10.1145/3689061.3689075. URL https://doi.org/10.1145/3689061.3689075.

Lung-Ming Chen, Yi-Hsiang Pan, and Yung-Jen Chen. A study of shuttlecock's trajectory in badminton. *Journal of Sports Science & Medicine*, 8(4):657–662, December 2009. ISSN 1303-2968. doi: jssm-08-657.

Yu-Jou Chen and Yu-Shuen Wang. Tracknetv3: Enhancing shuttlecock tracking with augmentations and trajectory rectification. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, MMAsia '23, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400702051. doi: 10.1145/3595916.3626370. URL https://doi.org/10.1145/3595916.3626370.

Jesse Davis, Lotte Bransen, Laurens Devos, Arne Jaspers, Wannes Meert, Pieter Robberechts, Jan Van Haaren, and Maaike Van Roy. Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned. *Machine Learning*, 113(9):6977–7010, September 2024. ISSN 1573-0565. doi: 10.1007/s10994-024-06585-0. URL https://doi.org/10.1007/s10994-024-06585-0.

Morten Holck Ertner, Sofus Schou Konglevoll, Magnus Ibh, and Stella Graßhof. Synthnet: Leveraging synthetic data for 3d trajectory estimation from monocular video. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, MMSports '24, pp. 51–58, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711985. doi: 10.1145/3689061.3689073. URL https://doi.org/10.1145/3689061.3689073.

Javier Galeano, Miguel Ángel Gomez, Fernando Rivas, and Javier M. Buldú. Entropy of badminton strike positions. *Entropy (Basel)*, 23(7):799, June 2021. ISSN 1099-4300. doi: 10.3390/e23070799.

Javier Galeano, Miguel Ángel Gómez, Fernando Rivas, and Javier M. Buldú. Using markov chains to identify player's performance in badminton. *Chaos, Solitons & Fractals*, 165:112828, 2022. ISSN 0960-0779. doi: https://doi.org/10.1016/j.chaos.2022.112828. URL https://www.sciencedirect.com/science/article/pii/S0960077922010074.

Jhen Hsieh. Neural network-based tracking and 3d reconstruction of baseball pitch trajectories from single-view 2d video. *arXiv preprint arXiv:2405.16296*, 2024.

Yu-Chuan Huang, I-No Liao, Ching-Hsuan Chen, Tsì-Uí Ík, and Wen-Chih Peng. Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8. IEEE, 2019.

Magnus Ibh, Stella Graßhof, and Dan Witzner Hansen. A stroke of genius: Predicting the next move in badminton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3376–3385, June 2024.

Paul Liu and Jui-Hsien Wang. Monotrack: Shuttle trajectory reconstruction from monocular badminton video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3513–3522, 2022.

Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 1520–1528, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.178. URL https://doi.org/10.1109/ICCV.2015.178.

Grand View Research. Sports analytics market size, share & trends analysis report by component (software, service), by analysis type (on-field, off-field), by sports (football, cricket, basketball, baseball), and segment forecasts, 2021-2028, 2021. URL https://www.grandviewresearch.com/industry-analysis/sports-analytics-market/. Accessed: [Insert access date].

Kanon Sato, Takuya Nakabayashi, Masahiro Yamaguchi, Kyota Higa, Ryo Fujiwara, and Hideo Saito. Time-consistent ball tracking and spin estimation with event camera. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, MMSports '24, pp. 59–64, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711985. doi: 10.1145/3689061.3689067. URL https://doi.org/10.1145/3689061.3689067.

Manoj Sharma, Monika, Naresh Kumar, and Pardeep Kumar. Badminton match outcome prediction model using naïve bayes and feature weighting technique. *Journal of Ambient Intelligence and Humanized Computing*, 12(8):8441–8455, August 2021. ISSN 1868-5145. doi: 10.1007/s12652-020-02578-8. URL https://doi.org/10.1007/s12652-020-02578-8.

Tomohiro Shimizu, Ryo Hachiuma, Hideo Saito, Takashi Yoshikawa, and Chonho Lee. Prediction of future shot direction using pose and position of tennis player. In *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pp. 59–66, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Baljinder Singh Bal and Gaurav Dureja. Hawk eye: A logical innovative technology use in sports for effective decision making. *Sport Science Review*, 21, 2012.

Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pp. 399–402, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930442. doi: 10.1145/1101149.1101236. URL https://doi.org/10.1145/1101149.1101236.

Nien-En Sun, Yu-Ching Lin, Shao-Ping Chuang, Tzu-Han Hsu, Dung-Ru Yu, Ho-Yi Chung, and Tsì-Uí Ík. Tracknetv2: Efficient shuttlecock tracking network. In *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, pp. 86–91, 2020. doi: 10.1109/ICPAI51961.2020.00023.

Gema Torres-Luque, Ángel Iván Fernández-García, Juan Carlos Blanca-Torres, Miran Kondric, and David Cabello-Manrique. Statistical differences in set analysis in badminton at the rio 2016 olympic games. *Frontiers in Psychology*, 10:731, April 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.00731.

Gema Torres-Luque, Juan Carlos Blanca-Torres, David Cabello-Manrique, and Miran Kondric. Statistical comparison of singles badminton matches at the london 2012 and rio de janeiro 2016 olympic games. *Journal of Human Kinetics*, 75:177–184, October 2020. ISSN 1640-5544. doi: 10.2478/hukin-2020-0046. Published by Sciendo.

Theresa Nkiru Uzor, David Chibuike Ikwuka, and Nonye Ann Ujuagu. Hawkeye technological innovation: Challenges and intervention strategies in sports. *J Mod Educ Res*, 2, 2023.

Shah Rutvik Vrajesh, A. N. Amudhan, A. Lijiya, and A. P. Sudheer. Shuttlecock detection and fall point prediction using neural networks. In *2020 International Conference for Emerging Technology (INCET)*, pp. 1–6, 2020. doi: 10.1109/INCET49848.2020.9154136.

Govind Waghmare, Sneha Borkar, Vishal Saley, Hemant Chinchore, and Shivraj Wabale. Badminton shuttlecock detection and prediction of trajectory using multiple 2 dimensional scanners. In *2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI)*, pp. 234–238, 2016. doi: 10.1109/CMI.2016.7413746.

Changfen Wang. Low-cost badminton trajectory recognition and landing point prediction optimization based on field coordinate system transformation. *Informatica*, 48(23), 2024.

Qiming Wang, Yongqiang Bai, and Hongxing Song. Middle fusion and multi-stage, multi-form prompts for robust rgb-t tracking. *Neurocomputing*, 596:127959, 2024a. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.127959. URL https://www.sciencedirect.com/science/article/pii/S0925231224007306.

Wei-Yao Wang, Hong-Han Shuai, Kai-Shiang Chang, and Wen-Chih Peng. Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton. In *AAAI Conference on Artificial Intelligence*, 2021. URL https://api.semanticscholar.org/CorpusID:244799733.

Wei-Yao Wang, Yung-Chang Huang, Tsi-Ui Ik, and Wen-Chih Peng. Shuttleset: A human-annotated stroke-level singles dataset for badminton tactical analysis. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 5126–5136, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599906. URL https://doi.org/10.1145/3580305.3599906.

Wei-Yao Wang, Wei-Wei Du, Wen-Chih Peng, and Tsi-Ui Ik. Benchmarking stroke forecasting with stroke-level badminton dataset. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024b. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/1042. URL https://doi.org/10.24963/ijcai.2024/1042.

Chris Beyer Wolf Gawin and Marko Seidler. A competition analysis of the single and double disciplines in world-class badminton. *International Journal of Performance Analysis in Sport*, 15 (3):997–1006, 2015. doi: 10.1080/24748668.2015.11868846. URL https://doi.org/10.1080/24748668.2015.11868846.

Erwin Wu and Hideki Koike. Futurepong: Real-time table tennis trajectory forecasting using pose prediction network. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:218482647.

Erwin Wu, Florian Perteneder, and Hideki Koike. Real-time table tennis forecasting system based on long short-term pose prediction network. In *SIGGRAPH Asia 2019 Posters*, pp. 1–2. 2019.

Momoe Yamane, Akimasa Kondo, Masashi Hatano, Ryosuke Hori, and Hideo Saito. Occlusion free multi-object tracking extention using multi-camera for sport. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, MMSports '24, pp. 1–6, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711985. doi: 10.1145/3689061.3689070. URL https://doi.org/10.1145/3689061.3689070.

Haoran Yi, Deepu Rajan, and Liang-Tien Chia. Automatic extraction of motion trajectories in compressed sports videos. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, pp. 312–315, New York, NY, USA, 2004. Association for

Computing Machinery. ISBN 1581138938. doi: 10.1145/1027527.1027599. URL https://doi.org/10.1145/1027527.1027599.

Zhengtao Zhang, De Xu, and Min Tan. Visual measurement and prediction of ball trajectory for table tennis robot. *IEEE Transactions on Instrumentation and Measurement*, 59(12):3195–3205, 2010. doi: 10.1109/TIM.2010.2047128.

A Appendix

Choice of Model Parameter To balance model capacity and computational efficiency, the dimensions of hidden space and final prediction layer are set to $d = 128$ and $d_F = 256$, respectively. And the auxiliary loss weight mentioned above is $\lambda = 0.3$. We chose 21 frames before the stroke to make predictions based on prediction accuracy. While using only 10 frames already yields a high accuracy of 91.4%, extending to 21 frames improves performance to 92.6%. Further increasing the frame count offers diminishing returns (e.g., 20 frames: 92.5%, 30 frames: 91.8%, 40 frames: 90.2%).

LLM Assistance

We used LLM to refine paper sections for clarity and grammar. We maintained full responsibility for reviewing and validating all LLM-assisted content, ensuring accuracy and scientific standards. LLM was not involved in core research, experimental design, data collection, or primary analysis. All scientific content, conclusions, and errors remain solely the our responsibility.