PRELUDE: ANTICIPATING LANDING POINTS AT RACKET-SHUTTLE CONTACT VIA CONTEXT-ADAPTIVE MULTIMODAL GATING

Anonymous authors
Paper under double-blind review

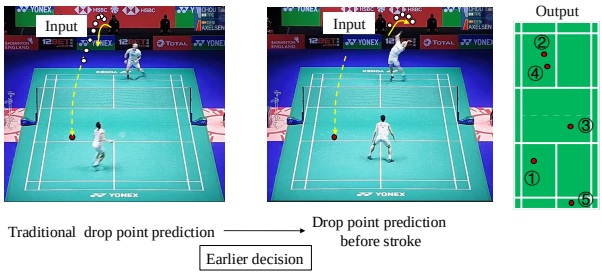

Figure 1: Traditional drop-point prediction analyzes the outgoing trajectory after contact. PRELUDE instead uses a fixed pre-contact observation window and returns a prediction 224 ms after detected contact. Physical robotic interception remains a future system-level validation.

1 ABSTRACT

We study pre-intervention anticipation under partial observability—predicting an outcome before the action completes. This setting requires synthesizing heterogeneous spatio-temporal signals and resolving inherent ambiguities in ground-truth labeling. We introduce PRELUDE, a conditional fusion framework for multimodal anticipation with a temporally bounded observation window. PRELUDE comprises three key elements: (i) a pre-contact multimodal encoder, (ii) a dual gating mechanism that dynamically reconciles cross-modal and spatio-temporal dependencies, and (iii) a physics-informed label generation protocol that maintains temporal separation. On standard benchmarks, PRELUDE sets a new state-of-the-art, achieving 92.6% accuracy while delivering an 85% latency reduction (224 ms end-to-end). By providing strong evidence that pre-stroke multimodal cues alone are sufficient for precise anticipation, this work opens a new direction for proactive analytics in competitive sports. Code is available at: https://anonymous.4open.science/r/AI-Sport-28D6/README.md.

2 INTRODUCTION

Consider a badminton player awaiting an opponent's smash. The shuttlecock is struck, and shortly afterward it lands on the player's side of the court. Predicting *where* it will land from observations ending at contact could provide earlier evidence for broadcast overlays, training analysis, and future robotic systems.

Yet most existing approaches to ball trajectory prediction do exactly the opposite: they wait until the ball has traveled for several frames after the stroke, then reconstruct or forecast its trajectory Liu & Wang (2022); Chen & Wang (2024). This is a *tracking* or *reconstruction* paradigm. It works well for post-hoc analysis, but it is inherently too late for real-time decision-making.

In this paper, we advocate for a fundamental shift: from *post-stroke reconstruction* to *pre-intervention anticipation*. We pose the following question: can we predict the shuttlecock's landing point at the moment of racket-shuttlecock contact, using only information available up to that instant? From a physics perspective, this is plausible—the trajectory is largely determined at impact, when the initial velocity, spin, and contact point are set. The challenge is to identify and integrate the right pre-stroke cues, and to do so fast enough to be useful.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We formalize a new problem setting: multimodal anticipation with a temporally bounded observation window. The prediction must be made at a pre-defined boundary—here, the moment of racket–shuttle contact—using only observations from a fixed past window. This setting is general and arises whenever an outcome is determined at an event but can only be observed beforehand. Badminton serves as a challenging instantiation because the relevant cues are heterogeneous, time-varying, and their importance depends on stroke context.

We identify four key pre-stroke modalities: (a) the 3D shuttlecock trajectory prior to contact, capturing incoming ball dynamics; (b) the player's arm gesture, encoding swing mechanics and intent; (c) the positions of both players, providing strategic context; and (d) the stroke type, serving as a high-level semantic prior. These modalities are heterogeneous, and their relevance depends on context—for a smash, the late trajectory matters most; for a drop shot, the early arc provides stronger cues. Static fusion cannot handle such variation.

To address this, we propose **PRELUDE**, a conditional fusion framework that dynamically modulates feature importance under a temporally bounded observation window. The framework introduces a dual gating mechanism that reconciles cross-modal and spatio-temporal dependencies, preventing dominant modalities from suppressing subtle but predictive cues. This design is motivated by the general challenge of integrating heterogeneous, partially observable signals at a pre-defined boundary.

We evaluate PRELUDE on TrackNetV2 and ShuttleSet22. It achieves 92.6% accuracy within 0.3 m and a measured 224 ms end-to-end latency. Ablations compare conditional gating with equal-input concatenation and standard attention. A FineGym component experiment is reported as a backbone stress test, not full-system cross-sport validation. In summary, our contributions are as follows.

- **Problem Formulation:** We formalize multimodal anticipation with a temporally bounded observation window, where predictions must be made at a pre-defined boundary using only past observations. This generalizes beyond sports to any setting where an outcome is determined at an event but must be anticipated beforehand.
- **Conditional Fusion Framework:** We propose PRELUDE, a conditional fusion framework that integrates heterogeneous modalities via hierarchical cross-attention and dual dynamic gating. The gating mechanism is designed to handle modality dominance and context-dependent relevance under partial observability.
- **Temporally Separated Data Protocol:** A physics-informed refinement protocol enforces temporal separation between input and label initialization, providing geometrically consistent supervision while reducing future information leakage.
- **Empirical Validation:** Controlled baselines, modality combinations, corruption tests, temporal shifts, and sensitivity studies evaluate accuracy, latency, and dependence on upstream extractors.

3 Related Work

3.1 3D Trajectory Prediction in Ball Sports

Recent advancements have leveraged high-precision equipment and deep learning to overcome these limitations in 3D Trajectory Prediction in Ball Sports. 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, paving the way for enhanced coaching methodologies and player performance improvements.

3.2 Multimodal Fusion in Machine Learning

Based on the point in the processing workflow where fusion occurs, multimodal fusion methods can be categorized into early fusion, middle fusion, and late fusion Gadzicki et al. (2020). Early fusion Barnum et al. (2020) combines data from different modalities before feature processing. This approach has the advantage of fully utilizing the information from various modalities, capturing inter-modal correlations, simplifying later processing, and often yielding more accurate results. However, early fusion can sometimes lead to a significant increase in feature dimensionality, thereby raising computational complexity, and may be adversely affected by the presence of low-quality modality information. Late fusion Snoek et al. (2005), on the other hand, involves processing each modality's data through the entire workflow individually, obtaining outputs from all unimodal models, and then aggregating these outputs to produce the final prediction. This method allows for the customized utilization of features from each modality but may fail to fully account for the relationships between modalities. Middle fusion Wang et al. (2024) occupies a position between early and late fusion, enabling flexible processing of features from each modality while balancing the integration of information from different modalities. However, its effectiveness often depends on the feature extraction methods and fusion strategies employed.

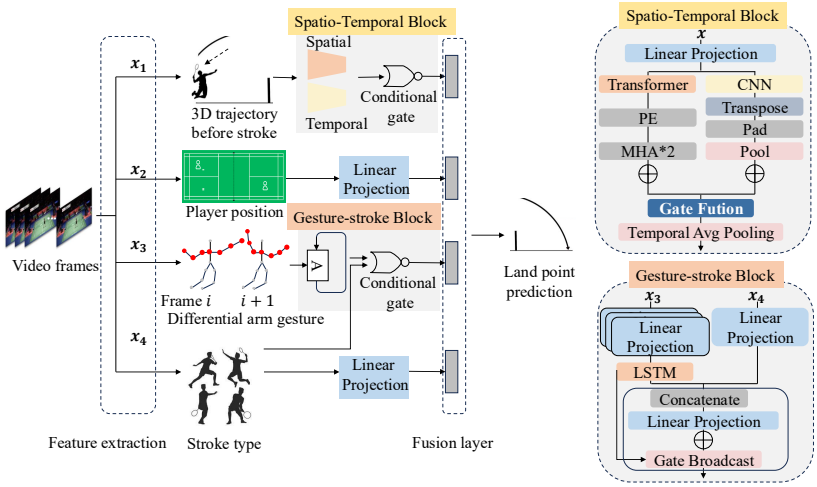

Figure 2: Illustration of the PRELUDE architecture. The model includes four inputs: the 3D trajectory of shuttlecock before stroke (X_1), two players' positions (X_2), the gesture feature (X_3, specifically 2D arm keypoints), and stroke type (X_4), with conditional gates put on X_1 and X_3 to enable dynamic feature recalibration, and outputs the shuttlecock landing point prediction.

4 Methodology

4.1 Overview

Let t_{impact} be the contact time detected by HitNet and let Δt denote the pre-contact window. Given T frames ending at $t_{\text{obs}} = t_{\text{impact}}$ (equivalently spanning $[t_{\text{impact}} - \Delta t, t_{\text{impact}}]$), the task is to predict the 2D landing point $\hat{y} \in \mathbb{R}^2$ without observing the outgoing flight. The prediction becomes available after feature extraction and inference; "contact-anchored" specifies the observation boundary rather than zero computation time. PRELUDE integrates four modalities: (1) 3D shuttlecock trajectory X_1 before stroke, (2) Spatio-temporal player coordinates X_2, (3) Player arm gesture features X_3, and (4) Stroke type classification X_4.

Formally, we define the prediction as a conditional function F:

$$\hat{y} = F(X_1, X_2, X_3, X_4; \Theta) \tag{1}$$

where Θ represents the learnable parameters. As illustrated in Figure 2, PRELUDE employs a hierarchical architecture featuring a **Cross-Fusion Mechanism** to preserve modality-specific characteristics, and **Conditional Gating** to dynamically recalibrate features under a temporally bounded observation window. The gating is conditioned on stroke context, allowing the model to adaptively weight spatial versus temporal evidence and to suppress modality-specific noise.

4.2 CROSS-FUSION MECHANISM

To facilitate deep interaction while preventing the "semantic drowning" of subtle cues (e.g., gestures) by dominant signals (e.g., trajectory), we propose a hierarchical cross-fusion strategy.

Let $F^{(1)} = [H_1, H_2, H_3, H_4]$ denote the intermediate features extracted from the first processing stage, and $F^{(2)} = [H'_1, H'_2, H'_3, H'_4]$ denote the refined features from the second stage. Instead of simple concatenation, we utilize a cross-attention mechanism to allow Stage 2 features to modulate the representations of Stage 1:

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

where the cross-attention operation is defined as:

$$\text{CrossAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

with query $Q = F^{(1)}W_Q$, key $K = F^{(2)}W_K$, and value $V = F^{(2)}W_V$. This design ensures that the model retains low-level details from $F^{(1)}$ while incorporating high-level semantic refinement from $F^{(2)}$, significantly outperforming standard late-fusion baselines.

4.3 SPATIO-TEMPORAL TRAJECTORY ENCODER

The 3D trajectory $X_1 \in \mathbb{R}^{3\times T}$ (where $T = 21$ frames) exhibits complex non-linear dynamics influenced by aerodynamics. We hypothesize that spatial geometry (kinematics) and temporal dynamics (rhythm) should be modeled separately before fusion.

Dual-Branch Transformation: We employ two parallel Transformers:

- *Temporal Branch:* A 2-layer Transformer encoder with Multi-Head Attention (MHA, 2 heads) to capture long-range temporal dependencies.
- *Spatial Branch:* A local windowed attention mechanism (window size=3) to model instantaneous velocity and acceleration patterns.

The transformations are defined as:

$$\mathbf{H}_1^{(t)} = \text{TemporalTransformer}(W_1X_1 + B_1) \tag{4}$$

$$\mathbf{H}_1^{(s)} = \text{SpatialTransformer}(W_1X_1 + B_1) \tag{5}$$

where W_1 and B_1 project the coordinates into a D-dimensional space. **Dynamic Conditional Fusion:** To adaptively balance spatial and temporal contributions, we introduce a learnable gating mechanism:

$$g_1 = \sigma(W_{g_1}[\mathbf{H}_1^{(t)} \oplus \mathbf{H}_1^{(s)}] + b_{g_1}) \tag{6}$$

$$\mathbf{H}_1 = \mu(g_1 \odot \mathbf{H}_1^{(t)} + (1 - g_1) \odot \mathbf{H}_1^{(s)}) \tag{7}$$

Here, σ denotes the Sigmoid function, μ denotes temporal averaging, and g_1 dynamically determines the weight ratio between the two domains.

4.4 CONDITIONAL GATE LSTM FOR GESTURE MODULATION

To model the dependency of landing point on player intention, we introduce a stroke-conditioned LSTM that filters irrelevant motion noise.

Gesture Feature Extraction: We extract 2D pixel coordinates of 6 key arm keypoints using OpenPose, forming $X_3 \in \mathbb{R}^{12\times(T-1)}$. First-order differencing is applied to capture motion flow.

Stroke-Conditioned Filtering: A standard LSTM processes the gesture sequence to obtain hidden states $\mathbf{h}$. However, to suppress noise from non-stroke movements (e.g., footwork), we condition the LSTM output on the stroke type embedding $\mathbf{H}_4 = \text{Embed}(X_4)$:

$$g_3 = \sigma(W_{g_3}[\mathbf{h}_{T-1} \oplus \mathbf{H}_4] + b_{g_3}) \tag{8}$$

$$\mathbf{H}_3 = \mathbf{h} \odot g_3 \tag{9}$$

This mechanism acts as a *contextual filter*: if the detected motion does not match the expected pattern for the predicted stroke type (e.g., a drop shot motion during a smash), the gate attenuates the signal.

Formalization as Contextual Modulation: We frame this interaction as a Feature Modulation problem. Let H_1 be the trajectory feature and H_3 be the gesture feature. The gate computes affine parameters:

$$(\gamma, \beta) = (\sigma(W_\gamma H_3 + b_\gamma), W_\beta H_3 + b_\beta) \tag{10}$$

The final modulated feature is:

$$\tilde{H}_1 = \gamma \odot H_1 + \beta \tag{11}$$

This formulation, analogous to Conditional Batch Normalization, allows the gesture context to both scale (γ) and shift (β) the trajectory representation, enabling richer interaction than simple concatenation.

4.5 Fusion and Prediction

The final fusion layer integrates features from all branches to regress the 2D landing point coordinates.

Hierarchical Aggregation:

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

Output Layer: The prediction $\hat{y}$ is generated by:

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

Loss Function: We optimize the model using Mean Absolute Error (L1 Loss) to ensure robustness against outliers in tracking noise:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{y}_i - y_i\|_1 \tag{14}$$

This formulation encourages the model to focus on precise spatial localization rather than being skewed by occasional large errors.

5 Experimental Study

5.1 Data Description

We combine TrackNetV2 Sun et al. (2020) and ShuttleSet22 Wang et al. (2023), containing 6,538 rallies and 13,582 candidate strokes. The reverse check discards 485 candidates. Each sample contains $T = 21$ pre-contact frames of shuttle coordinates ($X_1 \in \mathbb{R}^{3\times T}$), both players' court coordinates ($X_2 \in \mathbb{R}^{4\times T}$), first differences of six arm keypoints ($X_3 \in \mathbb{R}^{12\times(T-1)}$), a predicted stroke label X_4, and endpoint $y \in \mathbb{R}^2$. HitNet Liu & Wang (2022) supplies the contact anchor. The primary cross-dataset protocol uses four common stroke categories; a separate ShuttleSet22 experiment uses all ten categories. With ground-truth labels, the four- and ten-class settings obtain 93.53% and 94.31% landing accuracy; with predicted labels they obtain 92.60% and 88.64%, tracking classifier accuracy of 94.50% and 88.98%, respectively. We use an 8:1:1 train/validation/test split.

5.1.1 Evaluation Metrics

To evaluate the accuracy of landing 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 d is a predefined tolerance. We report MAE, MSE, accuracy at $d = 0.3$ m, and a threshold sweep from 0.10 to 0.30 m so the comparison does not depend on a single tolerance.

5.2 RALLY-CONSISTENT LANDING POINT GROUND TRUTH GENERATION

To obtain reliable landing-point labels without temporal leakage, we design a physics-informed geometric refinement protocol that leverages the true terminal landing of each rally as an offline consistency anchor. Model inputs are drawn exclusively from the 21-frame pre-contact interval, whereas label construction draws from a disjoint post-contact interval. For strokes with a visible terminal landing, the label is anchored directly to the observed court-plane intersection. For intercepted, non-terminal strokes, we recover a counterfactual endpoint via drag-based extrapolation constrained by court geometry, and then validate it against the rally-level terminal landing. This disjoint-interval design reduces direct frame leakage while providing geometrically consistent supervision, though it does not guarantee strict causality of upstream feature extractors.

Rally-level terminal landing detection. For each rally, we detect the true terminal landing position L_{rally} using TrackNet-based tracking followed by manual verification. This position is a real observation from the final stroke of the rally, not an extrapolation. It serves as an offline anchor for label validation and physics-parameter calibration.

Counterfactual anchor construction for non-terminal strokes. For the i-th stroke that is intercepted and does not directly land, we first obtain a candidate landing point $\hat{L}_i$ by extrapolating the observed post-contact trajectory segment (0 to 0.5 s after contact) under the standard drag model $m\dot{\mathbf{v}} = m\mathbf{g} - \frac{1}{2}\rho C_d A\|\mathbf{v}\|\mathbf{v}$ and intersecting it with the court plane. Instead of directly treating $\hat{L}_i$ as the label, we validate it against the rally-level terminal landing L_{rally}:

- If $\hat{L}_i$ and L_{rally} fall on the same half-court and their distance is below a tolerance (0.5 m in our implementation), we retain $\hat{L}_i$ as a medium-confidence counterfactual label.
- If the deviation is large, we discard the sample rather than force-fitting it, or optionally replace it with L_{rally} as a shared rally-level weak label. Because sharing a label across multiple strokes in the same rally reduces sample independence, we primarily rely on discarding and use the shared label only as a fallback.

Physics-parameter calibration. We use the rally-level true landings L_{rally} to calibrate the drag model parameters (e.g., C_d and air density) such that the extrapolated trajectories minimize error on known true landing points. The calibrated model is then used to re-extrapolate non-terminal strokes, yielding more reliable candidate anchors.

Confidence-aware supervision. We distinguish three confidence tiers during training:

- **High confidence:** terminal strokes with directly observed landings.
- **Medium confidence:** non-terminal strokes whose counterfactual anchors pass the rally-level consistency check.
- **Discarded:** non-terminal strokes whose anchors fail the consistency check.

The training loss is weighted by these confidence tiers, so that high-confidence samples dominate optimization while medium-confidence samples still contribute. Rally-level information is used only for offline label construction and never enters the model input, preserving the temporally bounded inference protocol.

Optimization details. The protocol proceeds in three steps. First, we extract X_1 from $t \in [-0.7\text{s}, 0]$ and initialize labels from observations at $t \in (0, 0.5\text{s}]$, ensuring the two windows share no frames. Second, we estimate a homography H that maps image-plane court intersections to the canonical court plane:

$$\min_{H} \sum_{i=1}^{N} \left\| y_i^{3D} - H^{-1} y_i^{2D} \right\|_2 . \tag{16}$$

Third, we fit a physics-constrained endpoint. MonoTrack uses a soft endpoint penalty,

$$\mathcal{L}_{\text{Mono}} = \sum_i \|\pi(\mathbf{Y}_i) - \mathbf{y}_i\|^2 + \lambda_{\text{plane}} Z_{\text{end}}^2, \tag{17}$$

Model	Acc	MSE	MAE
MonoTrack	80.5%	0.29	0.31
ShuttleNet-adapted	77.2%	0.33	0.34
DyMF-adapted	83.2%	0.28	0.30
RallyTemPose-adapted	84.8%	0.21	0.24
FCST-adapted	82.1%	0.29	0.32
BSTFormer-adapted	85.1%	0.22	0.25
ShuttleFlow-adapted	81.4%	0.27	0.29
Pre-stroke MLP	72.3%	0.38	0.40
Early Fusion (Concat)	79.6%	0.31	0.33
LSTM (no gate)	83.1%	0.25	0.27
Naive Concatenate	86.4%	0.19	0.21
PRELUDE	**92.6%**	**0.18**	**0.20**

Table 1: Controlled comparison under the contact-anchored protocol.

Method	0.10	0.15	0.20	0.25	0.30 m
MonoTrack	13.7	28.2	44.6	60.2	80.5
ShuttleNet	12.0	25.0	40.1	55.1	77.2
DyMF	18.9	36.0	56.3	73.0	83.2
RallyTemPose	17.4	37.6	53.5	69.7	84.8
FCST	18.0	35.8	54.3	73.0	82.1
BSTFormer	19.5	37.2	56.8	73.5	85.1
ShuttleFlow	16.8	34.5	52.1	68.9	81.4
Pre-stroke MLP	8.5	27.2	40.1	49.6	72.3
Early Fusion	14.2	27.8	42.3	58.9	79.6
LSTM (no gate)	19.5	37.1	55.8	71.8	83.1
Naive Con.	21.1	36.4	55.2	72.8	86.4
PRELUDE	**25.1**	**47.8**	**68.6**	**83.6**	**92.6**

Table 2: Accuracy (%) at increasingly strict distance thresholds.

whereas our refinement fixes the endpoint and estimates the parameter set $\theta = \{\mathbf{v}_0, \alpha, \beta\}$:

$$\min_{\theta} \sum_{t=1}^{T-1} \|\mathbf{Y}_{\text{sim}}(t;\theta) - \mathbf{Y}_{\text{obs}}(t)\|^2 \quad \text{s.t.} \quad \mathbf{Y}_{\text{sim}}(T) = \mathbf{Y}_{\text{land}}^{\text{anchor}}. \tag{18}$$

For terminal strokes, $\mathbf{Y}_{\text{land}}^{\text{anchor}}$ is the observed terminal landing. For non-terminal strokes, it is the rally-validated counterfactual anchor described above. During reverse checking, an independent annotator inspects the reconstructed trajectory overlaid on post-contact frames; parameters are never tuned in this check, and samples exceeding 0.5 m deviation are removed.

5.3 Landing Point Prediction Results Compared with Baseline Methods

We compare PRELUDE against MonoTrack Liu & Wang (2022), DyMF Chang et al. (2023), FCST Wang (2024), ShuttleNet Wang et al. (2022), ShuttleFlow Lien et al. (2025), RallyTemPose Ibh et al. (2024), and BSTFormer Chang (2026). Since these methods were originally designed for different output formats, we standardize them by replacing their task-specific head with an identical 2D regression head and training all models under the same data split and L1 objective.

Table 1 and Table 2 confirm that PRELUDE's advantage persists across all thresholds from 0.10 to 0.30 m. Notably, the gap widens at stricter thresholds: at 0.10 m, PRELUDE achieves 25.1% versus 21.1% for the next best method, indicating that conditional fusion is particularly effective for high-precision localization. Under equal input availability, PRELUDE improves accuracy over naive concatenation by 6.2 points. Among the two closest pre-stroke baselines, **BSTFormer** achieves 85.1% and **RallyTemPose** achieves 84.8%, both substantially below PRELUDE. This gap is informative: BSTFormer and RallyTemPose are designed for *stroke classification* or *next-stroke prediction*, and their architectures do not include explicit conditional gating or cross-modal recalibration. The three controlled baselines further isolate the source of PRELUDE's gain: (i) **Pre-stroke MLP** flattens all frames into a single vector and achieves only 72.3%, confirming that preserving temporal structure is essential. (ii) **Early Fusion** concatenates all modalities at the input layer but uses no intermediate cross-attention, reaching 79.6%; the 7.0-point gap to Naive Concatenation shows that hierarchical cross-fusion is necessary to prevent dominant signals from drowning out subtle cues. (iii) **LSTM (no gate)** replaces conditional gating with a standard LSTM, achieving 83.1% — a 9.5-point deficit relative to PRELUDE, directly quantifying the contribution of stroke-conditioned modulation. Table 2 confirms that the advantage persists across all thresholds from 0.10 to 0.30 m, rather than arising from a single favorable operating point.

Figure 3 visualizes the drop-point prediction performance alongside quantitative results. We randomly sample 500 test instances and plot the difference vectors between predicted and ground-truth drop points. PRELUDE concentrates errors near the origin, whereas the baselines exhibit broader spatial dispersion.

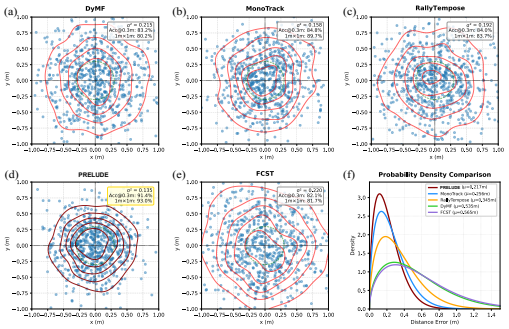

Figure 3: (a-e) Two-dimensional error distributions with KDE contours (red) and 0.3m threshold (green dashed). (f) Overlaid error-density histograms ($\alpha = 0.08$). PRELUDE has the strongest concentration near the origin.

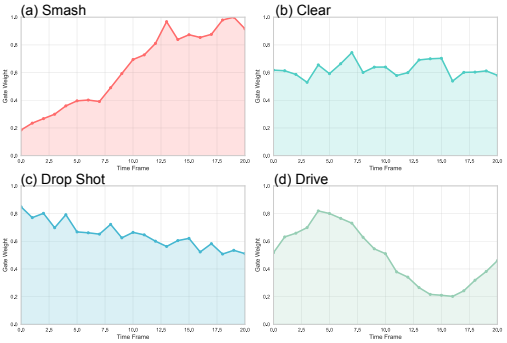

Figure 4: Average trajectory-gate weights by stroke type. The profiles differ across smash, clear, drop, and drive samples, consistent with context-dependent modulation.

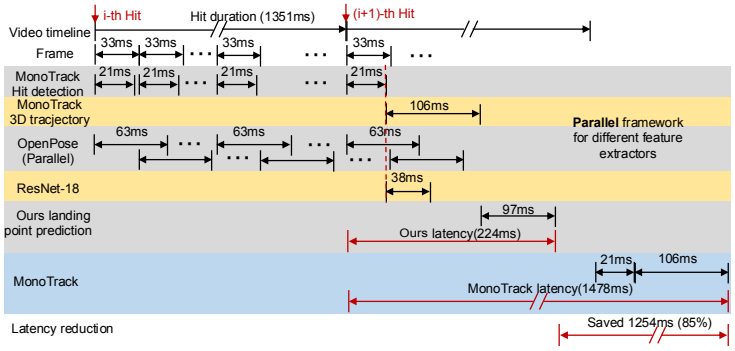

Figure 5: Real-Time Deployment Design and End-to-End Latency Breakdown.

5.3.1 Inference Time and Partial Observability

Latency was measured on an Intel i9-13900K with 64 GB RAM and a single RTX 4090. Hit detection and 3D reconstruction run sequentially, while OpenPose and ResNet-18 execute in parallel; the critical path is therefore $21 + max(106, 63, 38) + 97 = 224$ ms rather than the sum of every row in Figure 5. PRELUDE delivers an 85% latency reduction relative to MonoTrack's measured 1478 ms endpoint.

To characterize behavior under partial observability, Table 3 reports accuracy as a function of the number of pre-contact frames. PRELUDE retains 74.4% accuracy with only 5 frames and peaks broadly around 20–22 frames, indicating that the model degrades gracefully when observations are truncated. Table 4 further shows that shifting the contact anchor by up to 10 frames produces only a moderate accuracy drop, demonstrating robustness to imperfect contact detection.

5.4 Ablation Study on Feature Modalities

The modality study isolates the contribution of each of the four inputs. Trajectory (X_1) and gesture (X_3) are the strongest individual streams (64.5% and 64.6%), whereas stroke type alone reaches only 14.3%, confirming that static semantic labels cannot substitute for motion evidence. Combining the three visual streams yields 88.4%, and adding the predicted stroke type (X_4) lifts accuracy to 92.6%. This 4.2-point gain reveals a clear synergy: although X_4 alone is weak, its integration into a fully visual model acts as a semantic prior that disambiguates visually similar strokes (Table 5).

Frames	5	10	15	20	21	22	25
Naive concat.	57.3	68.4	76.1	84.2	86.4	86.1	84.2
Standard attn.	61.8	72.5	80.3	87.4	89.5	89.2	86.9
PRELUDE	**74.4**	**82.7**	**88.2**	**92.0**	**92.6**	**92.6**	**91.9**

Table 3: Fusion accuracy (%) versus available pre-contact frames.

Shift (frames)	0	2	4	6	8	10
Offset (s)	.000	.067	.133	.202	.267	.333
Accuracy (%)	92.6	92.0	91.4	89.0	88.5	87.4

Table 4: Sensitivity to contact-anchor shift.

Table 5: All non-empty combinations of shuttle trajectory (X_1), player position (X_2), gesture (X_3), and predicted stroke type (X_4). The table is laid out in two side-by-side columns to save vertical space.

X_1	X_2	X_3	X_4	Acc (%)	X_1	X_2	X_3	X_4	Acc (%)
–	–	–	–	00.0 ± 0.0	–	–	✓	✓	74.6 ± 1.9
✓	–	–	–	64.5 ± 0.9	✓	–	–	✓	76.7 ± 1.6
–	✓	–	–	23.0 ± 3.1	–	✓	–	✓	70.1 ± 1.9
–	–	✓	–	64.6 ± 1.3	✓	✓	✓	–	88.4 ± 0.2
–	–	–	✓	14.3 ± 2.6	✓	✓	–	✓	82.5 ± 1.0
✓	✓	–	–	74.8 ± 6.5	✓	–	✓	✓	86.8 ± 2.8
✓	–	✓	–	81.4 ± 2.0	–	✓	✓	✓	84.9 ± 1.0
–	✓	✓	–	85.9 ± 0.9	✓	✓	✓	✓	$\mathbf{92.6 \pm 0.2}$

5.4.1 Conditional Fusion via Gating Mechanism

To inspect how the model conditions on stroke type, we visualize the trajectory gate g_1 across stroke categories in Figure 4. The averaged temporal-branch weights differ systematically across smash, clear, drop, and drive samples, indicating that PRELUDE learns context-dependent modulation of spatial versus temporal evidence—consistent with the sport-physics intuition that different strokes rely on different temporal cues. Table 6 and Table 7 further show consistent gains over naive and standard-attention fusion across all stroke types.

Stroke	Naive	Attn	PRELUDE
Smash	89.2 ± 0.05	92.1 ± 0.05	93.2 ± 0.03
Clear	90.5 ± 0.05	93.4 ± 0.06	94.1 ± 0.04
Drop	81.6 ± 0.04	85.0 ± 0.05	88.5 ± 0.02
Drive	85.9 ± 0.04	88.8 ± 0.04	90.8 ± 0.03

Table 6: Fusion strategy by stroke type (accuracy, %).

Stroke	Acc	Stroke	Acc	Stroke	Acc
Smash	93.2	Clear	98.1	Drive	94.9
Lob	93.5	Drop	92.2	Push/Rush	91.7
Long serv.	91.3	Defensive	90.8	Net shot	90.9
Short serv.	88.5				

Table 7: Accuracy (%) of the 10-class model on fine-grained ShuttleSet22 labels.

6 Conclusion

We presented PRELUDE, a conditional fusion framework for multimodal anticipation with a temporally bounded observation window. We formalized the problem of predicting an outcome at a pre-defined boundary using only past observations, and proposed a dual gating mechanism that dynamically reconciles cross-modal and spatio-temporal dependencies. PRELUDE achieves 92.6% accuracy and a measured 224 ms end-to-end pipeline, compared with 86.4% for equal-input concatenation. Ablations and robustness tests support the framework on this benchmark while identifying upstream perception and counterfactual labels as the principal remaining limitations.

The problem setting may be applicable beyond badminton to domains where an outcome is determined at an event but must be anticipated beforehand, although full cross-domain validation remains future work.

REFERENCES

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.

Jing-Yuan Chang. Bst: Badminton stroke-type transformer for skeleton-based action recognition in racket sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 9889–9898, 2026.

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 AAAI Conference on Artificial Intelligence*, volume 37, pp. 6998–7005, 2023.

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*, pp. 41–49, 2024.

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.

Ranak Roy Chowdhury, Xiyuan Zhang, Jingbo Shang, Rajesh K Gupta, and Dezhi Hong. Tarnet: Task-aware reconstruction for time-series transformer. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 212–220, 2022.

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*, pp. 51–58, 2024.

Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I Webb, and Mahsa Salehi. Improving position encoding of transformers for multivariate time series classification. *arXiv preprint arXiv:2305.16642*, 2023.

Konrad Gadzicki, Razieh Khamsehashari, and Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pp. 1–6, 2020. doi: 10.23919/FUSION45008.2020.9190246.

Yongan Guo, Zhongyan Zhou, Yuao Wang, Na Zhu, Xuyun Zhang, Hongwang Xiao, Yuan Miao, and Bo Li. Rsfomer: Time series transformer for robust sports action recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9100–9109, 2025.

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 İk, 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*, pp. 3376–3385, 2024.

Yun-Hsuan Lien, Chia-Tung Lian, and Yu-Shuen Wang. Shuttleflow: learning the distribution of subsequent badminton shots using normalizing flows. *Machine Learning*, 114(2):39, 2025. doi: 10.1007/s10994-024-06682-0.

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 IEEE international conference on computer vision*, pp. 1520–1528, 2015.

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*, pp. 59–64, 2024.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2616–2625, 2020.

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

Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 399–402, 2005.

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.

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, 2024. 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, 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)*, 2023. URL https://api.semanticscholar.org/CorpusID:259262206.

Yu-Hsiang Wang et al. Shuttlenet: Position-aware fusion of rally progress and player styles for stroke forecasting in badminton. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

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*, pp. 1–6, 2024.

6.1 APPENDIX

6.2 Data Integrity and Leakage Analysis

To rule out spatial-temporal leakage through shared preprocessing, we introduce a manual verification test. On 200 samples with manually triangulated pre-contact X_1, MAE changes from 0.20 to 0.19 m and accuracy from 92.6% to 93.1% (Table 8). Paired tests detect no significant difference (MAE $p = .42$, MSE $p = .56$, accuracy $p = .40$), which is consistent with the interpretation that automatic tracking error is not the primary source of performance, although the small sample size and shared labeling pipeline prevent a definitive exclusion of circularity.

To test the generation of PRELUDE, we conduct experiment on dataset FineGym Shao et al. (2020) . FineGym lacks player position and stroke type, so we retain only the spatio-temporal encoder with OpenPose skeletons and replace the regression head with a classifier. This constitutes a component-level stress test that probes robustness under partial observability, extending the framework beyond its primary dataset. Results are shown in Table 9.

Feature source	MAE (m)	MSE	Acc (%)
MonoTrack	0.20	0.18	92.6
Manual multi-view	0.19	0.17	93.1

Table 8: MonoTrack versus manually triangulated pre-contact features.

Method	Accuracy (%)
RSFormer Guo et al. (2025)	89.5
TARNet Chowdhury et al. (2022)	87.2
ConvTran Foumani et al. (2023)	85.0
PRELUDE spatio-temporal component	**92.4**

Table 9: FineGym validation accuracy for the component-level adaptation.

AI Use Statement

Generative AI tools were used for language editing, LaTeX restructuring. The authors verified all technical claims, numerical results, equations, citations, and final text, and take responsibility for the submitted content.

Reproducibility Statement

The task protocol, model equations, metrics, data splits, implementation settings, latency decomposition, sensitivity analyses, and ablations are reported in the paper. An anonymized implementation is provided with the submission (Abstract).