**Comment 1**: The introduction has very few references. It would be surprising if there was not more relevant research in alternative approaches to table tennis analysis and/or ball tracking in other sports. This is needed to set your approach in context. Without this, it is difficult to assess whether the contribution of this paper has significance.

**Response 1**: Thank you for pointing this out. We agree with this comment. Therefore, we have expanded the introduction by including additional references to relevant research in table tennis analysis and ball tracking in other sports to provide better context for our approach. Specifically, we have added:

* A discussion on stroke detection and recognition in table tennis, referencing works like Kulkarni et al. (2023) and Voeikov et al. (2020), which leverage ball trajectory data and real-time video analysis.
* Examples of ball tracking methods in other sports, such as tennis and basketball, including studies like Huang et al. (2019) and Caio et al. (2023), which use deep learning for high-speed object tracking and 3D localization.

These additions highlight the novelty of our approach by situating it in the broader research context. The revised text can be found in the introduction section on page 1, paragraph 2, and lines 32–44.

**Comment 2**: What was the specific make and model of the "high-speed USB cameras" used? What makes them high-speed?

**Response 2**: Thank you for this question. We have clarified the specific make and model of the cameras used in our experimental setup. The cameras employed were two Logitech Brio 4K cameras, each operating at 90fps, and one Razer Kiyo Pro Ultra camera at 60fps. These cameras are considered "high-speed" due to their ability to capture video at high frame rates (90fps and 60fps, respectively), enabling detailed temporal resolution necessary for tracking fast-moving objects like table tennis balls.

This clarification has been added to the Materials and Methods section on page 2, paragraph 1, and lines 75–81.

**Comment 3:** Can OBS align video streams to less than the nearest whole frame (i.e., 16.7 ms at 60 FPS)? It would be good to evaluate what effect this could have on ball position estimations, e.g., how far can the ball travel during this time period?

**Response 3**: Thank you for raising this important question. To ensure precise alignment of video streams, we used the moment of ball contact with the table, captured by two cameras, as the synchronization reference point. This approach eliminates time misalignment between the videos by aligning the frames where the ball contacts the table.

When the videos are fully synchronized using this method, there is no temporal misalignment between the streams, meaning the ball position remains consistent across both views. As a result, there is no displacement error caused by timing discrepancies, and the issue of potential ball travel during a frame interval (16.7 ms at 60 FPS) is avoided entirely.

This clarification has been added to the Materials and Methods section in the revised manuscript to ensure the synchronization methodology is clearly explained.

**Comment 4:** Was motion blur a factor in any video captured during the experiments? If not, were there camera settings used to mitigate this?

**Response 4:** Thank you for this question. Motion blur was not a significant factor in the videos captured during our experiments. To mitigate motion blur, the cameras were configured with high shutter speeds of 1/500s or faster. These settings ensured that the ball's motion could be captured clearly even at high velocities. This detail has been added to the Materials and Methods section to clarify how motion blur was minimized during data collection.

This addition can be found on page 6, paragraph 《2.5. Ball Tracking》of the revised manuscript.

**Comment 5:** Was the modified YOLO11 network subjected to any additional training (e.g., transfer learning on your dataset)? Section 3.1 suggests this was the case but the full details are missing. Please provide relevant details, e.g., image resolution, number of images and how they were extracted from the videos and labelled, train/validation/test split, number of epochs training, etc. (The results show graphs for 300 epochs but don't specify if this was the full training time.)

**Response 5:** Thank you for pointing out the need for additional details regarding the training process for the modified YOLO11 network. We confirm that the YOLO11 network was subjected to transfer learning using our custom dataset. The following details have been added to the Materials and Methods section to clarify the training process:

1. **Image Resolution:** Training was performed with images resized to 640 × 640 pixels, which is the standard input resolution for YOLO.
2. **Dataset Details:** The dataset consisted of 2,000 images extracted from the experimental videos. These frames were carefully selected and manually labeled to include ball positions across various scenarios (serves, smashes, and rallies).
3. **Data Split:** The dataset was split into 70% for training, 20% for validation, and 10% for testing.
4. **Training Parameters:** The network was trained for 300 epochs in total, which corresponds to the full training time. A batch size of 16 was used, and the learning rate followed a cosine annealing schedule starting at 0.001.
5. **Hardware:** Training was conducted on an NVIDIA A100 GPU.

This additional information has been added on page 5, paragraph 《 **Ball Detection》** in the revised manuscript.

**Comment 6:** What was the prior art/related transformer models for the specific transformer architecture used in this paper? (The manuscript seems to reference papers for general transformer concepts rather than the specific architecture.)

**Response 6:** Thank you for your comment. We acknowledge that the manuscript referenced general transformer concepts but did not adequately detail the prior art specific to the transformer architecture used in our study. We employed a Transformer architecture for modeling the trajectory data, which is based on the standard Transformer encoder architecture introduced by Vaswani et al. (2017), with modifications tailored to our task.

The specific model used in our study, the **TrajectoryTransformer**, is a simplified version of the original transformer architecture, designed to handle sequential trajectory data with features such as the ball's 3D coordinates and frame indices. The architecture is characterized by the following:

* **Input Layer:** The input dimension was updated to 4 to account for the 3D coordinates (x, y, z) and frame index.
* **Transformer Encoder:** We utilized a standard Transformer encoder composed of stacked encoder layers, each with multi-head attention and position-wise feedforward networks, allowing the model to learn long-range dependencies in trajectory data.
* **Modifications:** The number of attention heads and the number of layers were reduced to suit the task of 3D trajectory prediction. The final output was passed through a classification layer to predict key points in the trajectory.

We have added a more comprehensive discussion of the prior work related to Transformer models for sequential data in the revised manuscript, including references to the application of Transformers in time-series forecasting and trajectory analysis.

This addition can be found on page 8, paragraph 《Transformer Model for Key Point Detection》.

**Comment 7:** How was the transformer model trained? What size was the dataset of labelled serves and the train/validation/test split and other relevant parameters of the training process? Did the dataset contain serves from multiple individuals (with different serve styles) and if so, how many different individuals?

**Response 7:** Thank you for your valuable comment. We have clarified the details regarding the training process of the transformer model and provided additional information on the dataset used for training. This addition can be found on page 8, paragraph 《Transformer Model for Key Point Detection.

**Comment 8:** In section 2.6, the identification of the serve sequence, the use of the threshold Y>0.5, etc. is not clear and needs clarification. Perhaps it refers to the coordinates in Figure 7b or Figure 12, but the reader should not have to guess.

**Response 8:** Thank you for pointing out the need for clarification. In Section 2.6, the **Y-coordinate** refers to the ball's position along the **forward-backward axis** relative to the table, with **Y = 0** set at the baseline on the player's side of the table. We use **Y = 0.5** as the threshold to segment the video and define the **end of the serve**. Specifically, this threshold corresponds to the point at which the ball crosses the boundary of the table and leaves the player's side.

We have now updated the manuscript to make this explanation clearer, specifying that **Y > 0.5** marks the point at which the serve action is considered complete. This serves as the critical boundary for cutting the video and identifying the serve's end.

These updates can be found in **Section 2.6** of the manuscript, where we clarify the role of the Y-coordinate in video segmentation.

**Comment 9:** Readers need to know details of the test sets that yielded the mAP values for ball detection, and F-score for trajectory key points. This is important to allow readers to evaluate how well the models were tested.

**Response 9:** Thank you for your valuable comment. We agree that explaining **mean Average Precision (mAP)** and F-score for trajectory key points

These clarifications have been added on **page 10, paragraph <** Result-**Ball Detection >, page 13,paragraph <** Result- **Transformer Model for Key Point Detection >,**  in the revised manuscript. We believe these details will help readers better assess the performance of our models.

**Comment 11:** F-score/F1-score is reported between 0 and 1. In the manuscript, a value of 93 is given. Was this F-score scaled by 100?

**Response 11:** Thank you for pointing this out. You are correct that the F-score is typically reported on a scale from 0 to 1. The value of **93** in the manuscript refers to an **F1-score of 0.93**, which was inadvertently written as 93 due to scaling by 100. We apologize for the confusion and have updated the manuscript to reflect the correct F1-score value of **0.93**.

**Comment 12:** There appears to be some discrepancy between the trajectories visualized in Figure 6, Figure 7, and Figure 9. The trajectories in Figures 6 and 7 are not smooth and appear to exhibit changing velocity and direction. It is less clear for Figure 9. The cause of the non-smooth trajectories needs further explanation since it suggests potential issues with ball detection and tracking.

**Response 12:** Thank you for your insightful comment. We acknowledge the discrepancy in the visualized trajectories in Figures 6, 7, and 9, and we understand the concern regarding the non-smooth trajectories.

The non-smooth trajectories observed in **Figures 6** and **7** are primarily due to a combination of two factors:

1. **Ball Detection Accuracy:** While the YOLO model performs well in general, occasional **false positives** or **missed detections** can occur, especially when the ball is moving rapidly or partially occluded by the player or other objects. These detection errors introduce small fluctuations in the trajectory, leading to abrupt changes in direction or velocity.
2. **Tracking Algorithm:** The **ByteTrack algorithm**, which is used for ball tracking, operates by associating detected frames with existing tracks. While it performs well under normal conditions, in cases of fast movements, occlusions, or brief detection failures, the track can briefly lose its continuity or incorrectly link to nearby detections, causing the trajectory to appear non-smooth.

For **Figure 9**, the tracking in this case is more stable, but there are still occasional moments where the ball's path appears less smooth due to minor occlusions or lighting conditions. We have clarified this in the revised manuscript, explaining that the occasional changes in velocity or direction are caused by temporary tracking inconsistencies or minor errors in ball detection.

We have added this explanation in the **Results** and **Discussion** sections to provide clarity on the causes of these non-smooth trajectories.

This addition can be found on **page 16, paragraph <Disscusion>, and lines 554–565** in the revised manuscript.

**Comment 13:** Certain sections of the paper feel padded and should be significantly reduced. In Section 2.5, Figure 5, and most of the description of ball tracking seems redundant. The authors used an off-the-shelf tracker without modification, so it seems inappropriate to describe the tracker in detail rather than just referring the reader to the original work for details.

**Response 13:** Thank you for your valuable feedback. We agree that the description of the ball tracking in Section 2.5, particularly the details in Figure 5, can be streamlined to avoid redundancy. Since the **ByteTrack algorithm** was used without modification, we have reduced the detailed description and instead referred the reader to the original work for further details.

We have also removed the excessive explanation of the tracker’s internal mechanics, focusing instead on its role and performance in our experiment. We have maintained the key points necessary to understand how the tracker contributed to the ball tracking process, without delving into unnecessary technical details.

These changes have been made in **Section 2.5** and are reflected in the revised manuscript. The updated text is more concise and focuses on the essential aspects of the ball tracking process.

**Comment 14:** In Section 3.1, Figure 8 does not seem to provide much value, and the text written in support of the figure makes claims that may not be well supported. I recommend deleting the figure and greatly reducing this section so as to focus on more important analysis suggested above.

**Response 14:** Thank you for your thoughtful suggestion. We agree that **Figure 8** does not add significant value to the overall analysis and that the claims made in its supporting text require further justification. As a result, we have removed **Figure 8** from the manuscript and substantially reduced the text in **Section 3.1**.

The section now focuses on the more critical analysis of the ball detection and trajectory key point detection performance, which is more relevant for the evaluation of the models. We have simplified the description to highlight the key findings and support them with clearer evidence from the results, avoiding any unsupported claims.

**Comment 15:** In Section 2.7, there does not seem to be any specific evidence provided to support the following claim: "The attention mechanism of the Transformer allowed it to discern subtle changes in the ball movement over time, capturing dependencies that traditional models might overlook." In the absence of evidence (which might require a comparative test against traditional models), the sentence should be removed.

**Response 15:** Thank you for pointing out the lack of explicit evidence to support our claim regarding the advantages of the Transformer model's attention mechanism. Due to space limitations, we could not include a full comparative analysis within the main text. However, we conducted detailed experiments comparing the Transformer model with MLP and LSTM using the same dataset. we have removed the sentence.

**Comment 16:** This paper focuses more on using existing methods (YOLO, Transformer, and Byte-Track) to solve real-world problems (Detecting serve fouls in table tennis), which appears to lack innovation. It is recommended that the authors clearly state the innovation and contribution of the paper.

**Response 16:** Thank you for your constructive feedback. We appreciate your observation and agree that it is important to clearly articulate the innovation and contribution of the paper, especially in terms of how existing methods are adapted and applied to solve the specific problem of detecting serve fouls in table tennis.

While the methods used (YOLO for ball detection, Byte-Track for tracking, and Transformer for key point detection) are established in the literature, the **novelty** of our work lies in their **integration and application** in the context of table tennis serve foul detection. This paper presents a **multi-camera, real-time system** that combines **3D trajectory analysis** and deep learning techniques to **automate the detection of serve fouls**, which is a **new approach** in the realm of sports officiating.

The key contributions of this paper include:

1. **Integration of multiple methods (YOLO, Byte-Track, and Transformer)** to address the challenges of serve foul detection in table tennis, particularly in real-time scenarios.
2. **Application of 3D trajectory analysis** to more accurately determine key points (throw point, highest point, hit point) and provide **objective foul detection** based on established rules.
3. The **novelty of combining the models** in a unified system, providing a more effective and scalable solution than existing manual or semi-automated methods.
4. **Real-world applicability**, showing how these techniques can be used to support referees in live settings, enhancing fairness and consistency in table tennis competitions.

**Comment 17:** This method includes many steps, such as frame synchronization, intrinsic and extrinsic calibration, ball detection, ball tracking, and video segmentation. Most of them are qualitative demonstrations and lack quantitative evaluation, which makes it difficult for readers to determine the final effect that the combination of these steps can achieve. I suggest that the authors should provide a quantitative evaluation of each step, which will help readers understand which part played a key role.

**Response 17:** Thank you for your insightful feedback. We understand the importance of providing quantitative evaluations for each step of the process to help readers assess the contribution of each component to the overall performance of the system.

**Frame Synchronization:** Updated in Section 2.2 (Data Collection and Synchronization) to clarify that complete synchronization of video streams is achieved using the key impact moment of the ping pong ball.

**Intrinsic and Extrinsic Calibration:** Detailed comparisons in the experiment indicate that error decreases as the distance to the calibration markers decreases. To minimize error, we used a calibration setup covering a wide area around the ping pong table. Although the experimental comparisons are extensive, the page limit prevents us from fully elaborating on them here.

**Ball Detection:** Quantitatively evaluated using mean Average Precision (mAP).

**Video Segmentation:** Achieved 100% accuracy by leveraging precise 3D coordinates for segmentation.

**Comment 18:** Some of the images in this article are not very clear, i.e., Figure 2 and Figure 7. It is recommended that the authors try to replace the images with vector graphics.

**Response 18:** Thank you for your constructive feedback. We appreciate your observation regarding the clarity of certain images. We have reviewed **Figures 2** and **7**, and we agree that they could be clearer. In response, we have replaced these images with higher-resolution versions and converted them to **vector graphics** to ensure better clarity and scalability, especially for publications where image quality is critical.

These updated images will now appear much sharper, even when zoomed in or printed at larger sizes. The revised versions of **Figures 2** and **7** are included in the updated manuscript.

**Comment 19:** In the Section of Results, the authors should specify whether the methods at each stage need to be retrained, how the dataset was created, how large the dataset is, how some necessary hyperparameters were set, etc., so that other researchers can reproduce the results.

**Response 19:** Thank you for your helpful comment. We have added detailed explanations in the Materials and Methods section under Experimental Setup, Ball Detection, and Transformer Model for Key Point Detection. These additions specify whether methods at each stage require retraining, describe the dataset creation process, its size, and provide information on key hyperparameters to ensure the reproducibility of our results.

**Comment 20:** The authors should consider adding a brief literature review about existing object detection approaches, based on single and multi-cameras, along with the applicable object detection models. The review can help support the authors’ system design as aligned with the state-of-the-art of small object detection.

**Response 20:** Thank you for this insightful suggestion. We agree that a brief literature review on existing object detection approaches, especially for small object detection using single and multi-camera setups, would provide valuable context for our work. To address this, we have added a section in the **Introduction** where we discuss key approaches in object detection, including both single-camera and multi-camera systems, and how they relate to small object detection.

**Comment 21:** The authors are recommended to have a review of related studies in object detection for table tennis and highlight how the proposed study contributes to this field of research.

**Response 21:** Thank you for your valuable suggestion. We agree that a review of related studies in **object detection for table tennis** would provide important context for our work and demonstrate how our study contributes to this field. We have now added a section in the **Introduction** that specifically reviews existing work in table tennis object detection and discusses the gap that our study addresses.

**Comment 22:** In Sections 3.2 to 3.4, it is recommended to assess the performance of ball tracking and key point detection with quantitative metrics to support the effectiveness of the proposed approach.

**Response 22:** Thank you for this suggestion. We agree that providing **quantitative metrics** for ball tracking and key point detection will offer a clearer assessment of the performance of our approach. In response, we have now included several relevant metrics in **Sections 3.2, 3.3, and 3.4** to quantitatively evaluate the effectiveness of the **ball tracking** and **key point detection** components.