Video-Based Tennis Biomechanics: Automated Injury Risk and 8-Stage Assessment

Anika Arugunta
University of Washington
arugani9@uw.edu

Eun Be
University of Washington
eunbe@uw.edu

Kevin Wu
University of Washington
kwu2004@uw.edu

Alexander Metzger
University of Washington
metzgera@uw.edu

Abstract

We present a marker-less video pipeline for fine-grained tennis serve analysis from ordinary monocular video. Our system first segments a serve into the eight biomechanical stages described by Kovacs and Ellenbecker, then evaluates stage-specific pose, racket, and ball measurements with a rule-based technique and injury-risk analyzer. We collected and annotated a recreational tennis serve dataset with exact stage boundaries, compared several temporal action segmentation (TAS) architectures and feature families, and integrated the best segmenter with a diagnostic reporting tool. The strongest configuration, an MS-TCN++ model using pose/racket features, ASRF-style boundary regression, and constrained Viterbi decoding, reached 91.8% frame accuracy and 93.2 segmental F1 at IoU 0.50 on the test split. A case study on a held-out serve shows how the segmented stages support interpretable frame-level feedback, including insufficient knee bend warnings, passing contact-height checks, and follow-through risks.

1 Introduction

The tennis serve is a highly complex, high-velocity movement requiring precise whole-body coordination. Because of the repetitive joint stresses involved, improper technique compromises competitive performance and elevates the risk of musculoskeletal injuries to the shoulder, elbow, and spine. Thus, players and coaches have a strong interest in capturing biomechanical data to optimize technique and prevent injury.

In professional tennis, detailed biomechanical feedback often relies on multi-camera motion capture systems or expert coaching. Commercial mobile applications offer more accessible alternatives, but they remain closed systems and generally do not expose the frame-level mechanics needed for reproducible research. Conversely, existing open-source solutions are often restricted to coarse-grained video classification, manual comparison against reference videos, marker-based motion capture, or summary statistics rather than phase-specific feedback. A significant gap remains for an accessible, end-to-end framework that delivers granular, serve analysis from standard, single-view video.

To address these limitations, we present a novel, marker-less video analysis pipeline for automated tennis serve biomechanics and injury risk assessment. Our primary contributions are twofold:

- **A Temporal Action Segmentation Dataset of Recreational Tennis Serves:** We introduce a curated video dataset of tennis serves capturing diverse techniques representative of recreational tennis players. Each video is hand-labeled to mark the exact temporal boundaries of the eight distinct serve stages.

- **An End-to-End Marker-less Injury Risk Analysis Pipeline:** We develop an automated pipeline to evaluate tennis mechanics. The pipeline leverages spatial and spatiotemporal feature extraction and Temporal Action Segmentation (TAS) to map video frames to the eight stages of a tennis serve. It then combines the predicted TAS segmentations with spatial pose data to drive an evaluation engine governed by 19 biomechanical rules, automatically generating diagnostic reports on joint angles, heights, and injury risks.

All authors contributed equally with direct collaboration on data collection and annotation. Anika then took the lead on consolidating our separate annotation formats and sampling the train-test split while Kevin compared pose models. Afterwards, Eun Be constructed the biomechanical rules model for risk detection and Alex took the lead on developing the TAS model.

2 Related Works

Our stage definitions follow the tennis-specific eight-stage serve model of [6], which decomposes the serve into preparation, acceleration, and follow-through phases and argues that a tennis-specific model is more appropriate than generic overhead-throwing models. This framing is central to our system: instead of asking whether a video contains a serve, we ask which biomechanical stage each frame belongs to and evaluate each rule only within the relevant stage.

Pose estimation provides the spatial substrate for this analysis. We evaluated several marker-less pose estimators qualitatively and selected YOLO11-pose for the draft TAS features because it supports human keypoint estimation and efficient deployment in the Ultralytics framework [4]. We also evaluated YOLO26-pose, the newer Ultralytics model family that includes pose-specific variants and reports stronger keypoint performance [5]. We also incorporate racket and ball features from RacketVision, which defines a racket-sports benchmark for fine-grained ball tracking and articulated racket pose estimation [2]. These explicit sports-object features are important because racket and ball positions are often decisive at the contact and follow-through stages.

Temporal Action Segmentation (TAS) predicts one action label per video frame rather than a single label for the entire clip. We compare three representative TAS families. MS-TCN++ uses multi-stage dilated temporal convolutions to generate and refine framewise predictions, addressing over-segmentation through progressive refinement [7]. ASFormer adapts Transformer-style sequence modeling to action segmentation while retaining locality and hierarchical structure [10]. FACT models frame and action tokens jointly with cross-attention for efficient action segmentation [8]. We further adapt the Action Segment Refinement Framework (ASRF), which predicts action boundaries to alleviate over-segmentation [3]. Finally, we compare explicit pose/racket features against RGB video embeddings from I3D [1] and VideoMAE V2 [9].

3 Methods

Our method combines a manually annotated two-camera serve dataset, qualitative pose-model selection, configurable temporal action segmentation, and a stage-conditioned biomechanical rule engine.

3.1 Dataset Collection and Annotation

Our dataset consists of approximately one hour of tennis serve videos captured at an outdoor recreational court. The videos were recorded using two cameras: one positioned to capture a full body view of the serving player and the second positioned to see where the ball lands for future labeling of the serve outcomes. As shown in Figure 1, both phones were mounted on a music stand to provide a simple, reproducible court-side capture setup. The serves span a mix of both successful and unsuccessful attempts. Existing tennis serve datasets were not used because they lack temporal stage segmentation labels and consist primarily of professional players, making them less representative of recreational player mechanics – which is the main target audience of our system since professional players have access to advanced hardware and coaching.

The videos were then manually clipped and annotated, noting the representative frame ranges for each stage of the serve. The eight stages of a serve consisted of: (1) Start, (2) Release, (3) Loading,

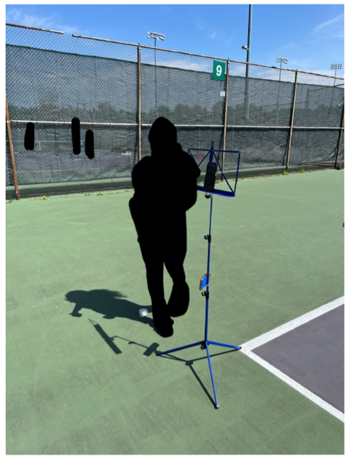


Figure 1: Data collection setup. A music stand held one phone facing the tennis player for full-body serve analysis and a second phone facing the court to record the serve outcome.

(4) Cocking, (5) Acceleration, (6) Contact, (7) Deceleration, and (8) Finish. Finally, the serve clips were split into 80% training and 20% test sets for model evaluation.

3.2 Pose Embedding Evaluation

A foundational component of our biomechanical analysis pipeline is the extraction of precise 2D human pose data across all video frames. Because our custom dataset consists of unconstrained, single-view videos without optical marker or multi-camera motion capture ground truth, direct quantitative metric evaluation (such as Mean Per Joint Position Error) was unfeasible.

To determine the most robust architecture for our use case, we conducted an empirical qualitative evaluation across four state-of-the-art human pose estimation models: YOLO11, YOLO26, MediaPipe Pose, and ViTPose. Each architecture was evaluated across a subset of videos using a qualitative grading rubric. This rubric assessed three performance dimensions: (1) keypoint localization accuracy, (2) tracking stability and robustness against motion blur, and (3) temporal consistency in limb-side identification (left-to-right joint attribution) during torso rotation and self-occlusion.

3.3 Temporal Action Segmentation

Figure 2 summarizes the TAS pipeline. Each input serve clip is processed into two complementary feature families. The first family consists of explicit structured features: YOLO11 2D human pose keypoints, RacketVision racket keypoints, ball position, normalized joint coordinates, velocities, joint angles, angular velocities, and derived racket/ball quantities such as racket-head speed and ball-to- racket distance. The second family consists of learned RGB embeddings from I3D and VideoMAE V2. We train TAS models to map these framewise features to the ordered label sequence

$$\text{Start} \rightarrow \text{Release} \rightarrow \text{Loading} \rightarrow \text{Cocking} \rightarrow \text{Acceleration} \rightarrow \text{Contact} \rightarrow \text{Deceleration} \rightarrow \text{Finish.}$$

We implemented a YAML-driven experiment system so that each component can be ablated independently. The model family, feature sources, boundary-head targets, and decoding strategy are all specified in the experiment configuration. The boundary branch is inspired by ASRF and is trained on selected transition targets. Our main configuration applies boundary supervision to contact-start, because the contact stage is only one or two frames in many clips and tends to dominate downstream biomechanical timing. We also experimented with a broader boundary set consisting of release-end, loading-end, cocking-end, contact-start, contact-end, and finish-start.

At inference time, optional constrained decoding enforces the known monotonic structure of a tennis serve. Given per-frame stage scores, the dynamic program permits only self-transitions or transitions from stage i to stage $i+1$, and it requires exactly one nonempty segment for each of the eight stages. When boundary scores are available, they are used as transition evidence. This decoder prevents

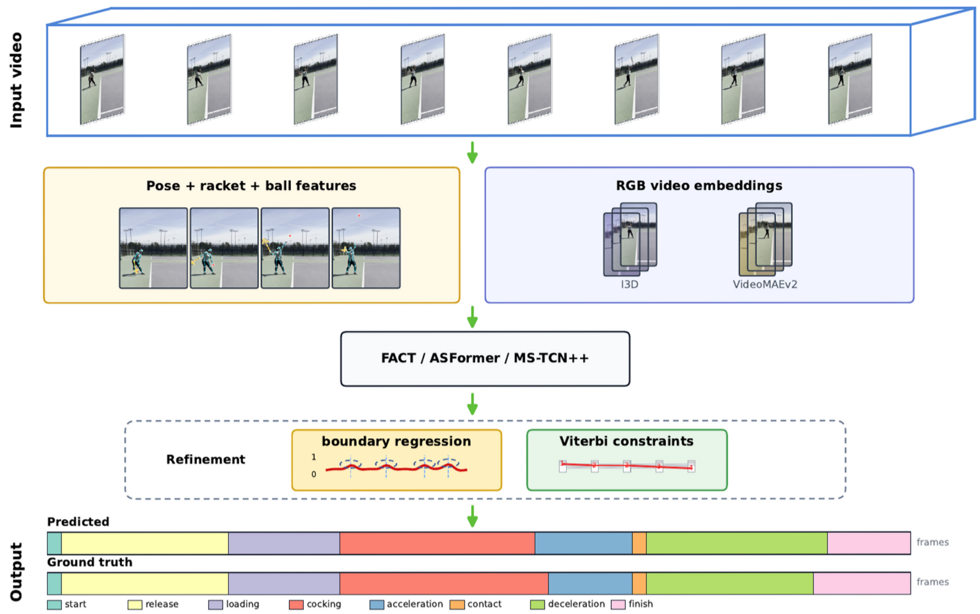


Figure 2: Overview of our temporal action segmentation pipeline. Input serve clips are converted into explicit pose/racket/ball features and optional RGB video embeddings, passed through one TAS backbone (FACT, ASFormer, or MS-TCN++), optionally refined with boundary regression and constrained decoding, and finally compared against ground-truth eight-stage labels.

impossible sequences such as returning from follow-through to loading and makes the model better aligned with the biomechanical prior.

3.4 Injury Risk and Technique Analysis

In order to analyze the serving technique of the player and flag any potential injury risks, we developed a rule-based biomechanical evaluation tool to automatically assess each serve. Using the keypoints from the pose extractions and the representative frame range for each stages of the serve as an input, the tool runs 19 biomechanical checks distributed across all eight serve stages.

Each rule computes geometric quantities such as joint angles, body ratios, and relative heights using the coordinates of the keypoints. The rules were derived from the 8-stage tennis serve model proposed by [6]. For instance, at the contact stage, the model indicates that there needs to be a slight elbow flexion; therefore, the tool checks for degree of arm extension during that stage. Overall, the tool checks many different factors, including the degree of leg bend, wrist positioning in relation to the elbow, and shoulder alignment based on the 8-stage model.

The exact rule list is provided in Appendix A. For each rule, the script selects the most relevant frame in the predicted stage, draws the pose, cleaned racket keypoints, and cleaned ball annotation, and then overlays the measured limb, joint angle, height, ratio, or midline used by the rule. The overlay background is green for passed rules, yellow for warnings, and red for risks. Appendix B lists the full report for the representative held-out serve shown in Figure 6.

We assessed this tool qualitatively by examining whether its outputs on representative serves align with expected biomechanical patterns, and whether flagged risks correspond to visually observable deviations in the video.

4 Results

The results show that explicit pose/racket features with MS-TCN++, boundary supervision, and constrained decoding produced the strongest stage segmentation, while the downstream rule engine generated readable frame-grounded diagnostics for held-out serves.

4.1 Pose Embedding Evaluation

The qualitative evaluation revealed pronounced differences in robustness across the four pose estimation models when run on our dataset.

YOLO11 emerged as the optimal framework for our pipeline [4], demonstrating superior keypoint localization accuracy and temporal consistency. It consistently locked onto high-velocity anatomical keypoints, such as the wrist and ankle, even during periods of significant motion blur at the acceleration and contact stages. Furthermore, YOLO11 maintained complete limb tracking continuity through complex body rotations without experiencing identity switches.

Conversely, the remaining models suffered from recurrent failure modes:

- **YOLO26:** Though capable of localized keypoint accuracy and designed as the newer Ultralytics pose family [5], YOLO26 exhibited tracking inversions midway through the serving motion on our clips. Specifically, during stages with torso rotation and self-occlusion such as the acceleration stage, the model routinely swapped left and right arms, mislabeling the right arm swinging the racket as the left arm. Such identity flipping injects fatal errors into downstream biomechanical computations.
- **MediaPipe Pose:** MediaPipe executed poorly on our dataset, generating highly unstable and noisy keypoint coordinates. The model proved largely incapable of handling recreational tennis serves, introducing significant inaccuracy and erratic coordinate jitter.
- **ViTPose:** ViTPose produced reasonably smooth and stable tracks but sometimes struggled with spatial accuracy. The predicted skeletons often failed to align completely with the actual contours of the player's physical form, yielding structural offsets that warp the absolute angle and height measurements required by our rule engine.

These failure modes are illustrated in Figure 3. Based on these qualitative observations, YOLO11 was selected as the sole pose estimation model to drive the temporal action segmentation and injury risk assessment steps of our pipeline.

YOLO26 limb swap   MediaPipe distortion   ViTPose body offset

Figure 3: Representative pose estimation failure cases for the non-YOLO11 models. YOLO26 produced limb-side inversions during torso rotation, MediaPipe generated unstable distorted skeletons, and ViTPose produced smoother but visibly misaligned body estimates.

4.2 Temporal Action Segmentation

Table 1 reports representative TAS results. The best performing configuration was MS-TCN++ with pose+racket features, contact-start ASRF boundary supervision, and constrained Viterbi decoding. This model achieved 91.8% frame accuracy and 93.2 segmental F1 at IoU 0.50. The result suggests that explicit geometric features are well matched to the short, highly structured serve clips, while boundary supervision is especially useful for the brief contact transition.

Experiment	Frame Acc.	F1@50	Avg. Stage Acc.
MS-TCN++ contact	**91.8**	**93.2**	**87.8**
MS-TCN++ ASRF	91.4	88.1	84.5
MS-TCN++ Viterbi	91.1	86.4	84.6
MS-TCN++ pose only	90.4	86.9	83.3
ASFormer	89.6	86.4	83.0
FACT	65.2	22.2	35.7
MS-TCN++ VideoMAE	91.5	87.5	84.0
MS-TCN++ I3D	81.1	54.0	61.7

Table 1: Temporal action segmentation results on the test split. Metrics are percentages. "MS-TCN++ contact" denotes the pose+racket model with contact-start boundary supervision and constrained Viterbi decoding.

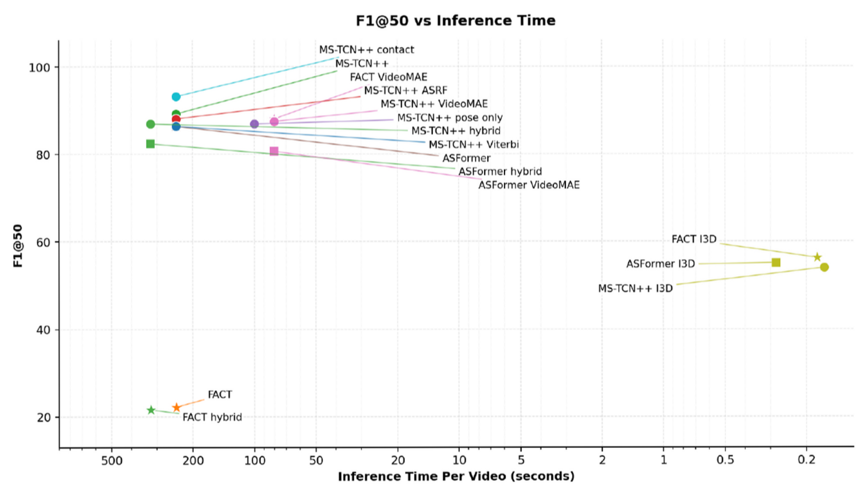
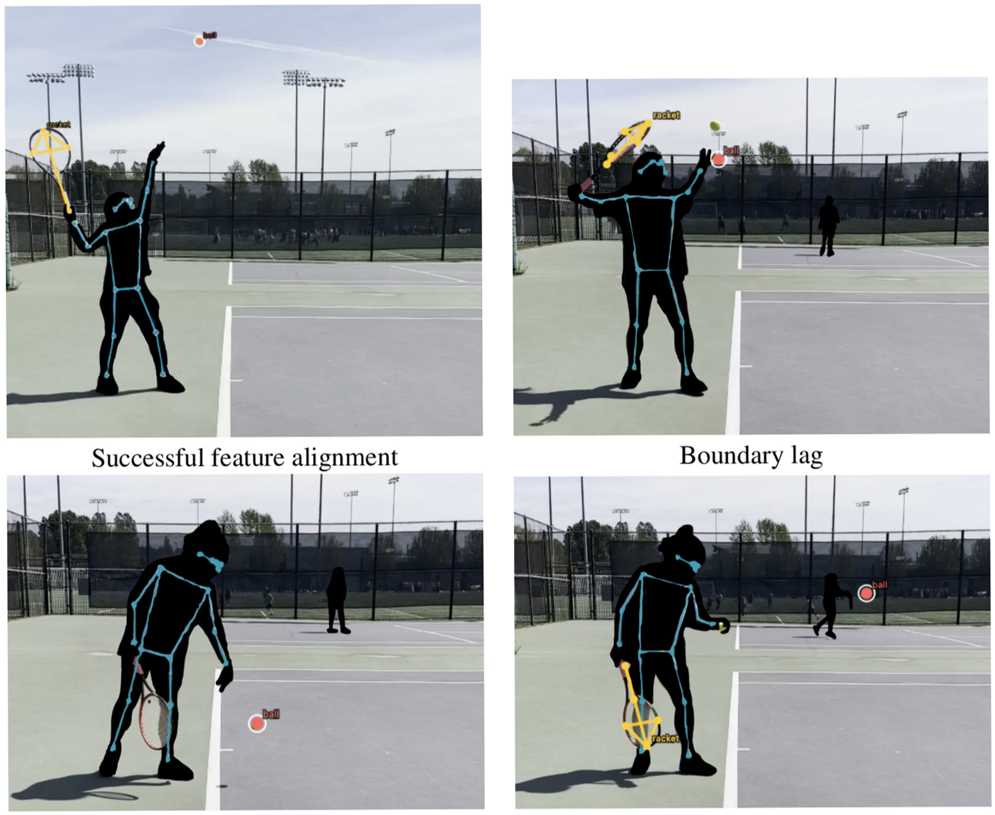


Figure 4: Accuracy-speed tradeoff for TAS experiments. The horizontal axis includes feature extraction time plus model inference and decoding time. Explicit pose/racket models are slower because they include pose and racket/ball extraction, but the best MS-TCN++ configuration gives the strongest segmental accuracy.

Qualitatively, the main TAS failure modes were boundary lag around contact, missing or noisy racket/ball cues, and confusion when the player remains nearly static before or after the serve motion. These errors are visible in Figure 5. The most important lesson from the ablations is that implicit RGB embeddings alone can be surprisingly competitive for coarse frame accuracy, but they are less reliable for exact boundary placement. For downstream biomechanics, a one-frame contact error can move the rule engine to the wrong physical moment, so we prioritize boundary-sensitive metrics and constrained decoding rather than frame accuracy alone.

4.3 Injury Risk and Technique Analysis

We ran the injury risk analysis tool on a representative sample of serves from the dataset, covering both visibly well-executed and faulty serves. Across all runs, the tool was able to successfully produce reasonable diagnostic reports for each of the serves.



Successful feature alignment

Boundary lag

Missing racket cue

Background/static confusion

Figure 5: Representative TAS qualitative examples. Explicit features are most reliable when pose, racket, and ball cues agree, but brief stage boundaries, missing racket detections, and misleading detections from players in the background remain challenging.

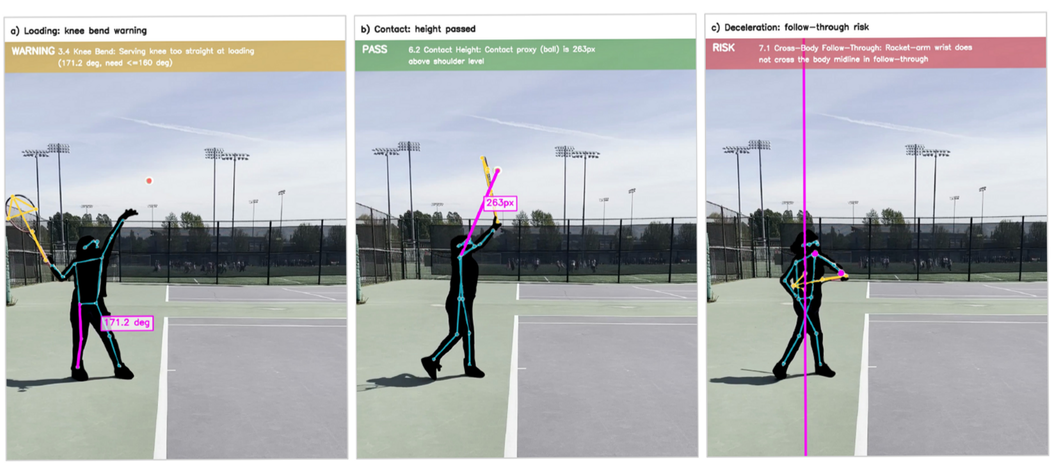


Figure 6: Representative rule outputs for held-out test serve 5. The analyzer overlays YOLO11 pose, cleaned racket keypoints, cleaned ball annotations, and the specific measurement used by each rule. Yellow indicates a warning, green indicates a pass, and red indicates a risk.

Qualitative evaluation of the outputs of the tool indicates that the tool's results are consistent with visual assessment of the serve footage. For instance, on serves with visible issues such as overly bent arms in the contact stage, the report flagged that rule as a risk, showing the degree it measured versus what is required. Likewise, for serves that looked visually strong, the report indicated a high number of passing rules with a few warnings concentrated around stages involving rapid movement where keypoint confidence was lower. The tool strongly relies on the quality of the pose estimates and representation frames for accurate results.

Overall, the tool demonstrated that automated, stage specific biomechanical injury risk analysis is feasible from a single view monocular video without expert manual review. However, it should be noted that its diagnostic precision remains bounded by the geometric limitations of 2D pose estimation.

5 Discussion

The strongest part of the system is the coupling between stage-aware segmentation and interpretable geometric rules. TAS converts a raw serve clip into a small number of biomechanically meaningful windows, allowing the rule engine to ask targeted questions: whether the tossing wrist is above the tossing shoulder during release, whether the serving knee is flexed during loading, whether the arm is extended at contact, and whether the follow-through crosses the body midline. This makes the outputs more actionable than an end-to-end classifier that only predicts "good" or "bad" technique.

The experiments also show why sports biomechanics is a hard setting for TAS. Several stages last only a handful of frames, and contact may last a single frame at 30 fps. Models that appear reasonable under frame accuracy can still fail at the precise instant needed for the downstream rule. The constrained Viterbi decoder helps by enforcing the known serve-stage order, but it cannot recover from severely wrong feature evidence. Racket and ball keypoints are particularly fragile because the ball is small and the racket is often motion-blurred or partially occluded by the player.

5.1 Limitations and Future Work

While our end-to-end pipeline effectively automates temporal stage segmentation and injury risk profiling, several limitations remain that point to opportunities for future work. First, our dataset is bounded in scope, consisting of a relatively small cohort of recreational players captured from a uniform camera angle. Expanding the dataset to incorporate a larger, multi-institutional subject pool with wider demographic diversity would greatly improve the generalization and robustness of the underlying TAS backbones. Furthermore, capturing variations across multiple heterogeneous camera angles (e.g., posterior, lateral, and oblique views) is necessary to make the system fully robust to varying real-world court conditions and filming constraints.

Second, our current pipeline relies entirely on 2D spatial coordinates extracted from monocular, single-view videos. While 2D pose tracking provides an affordable framework for consumer mobile deployment, it introduces inherent geometric limitations, such as foreshortening artifacts and out-of-plane joint self-occlusions during intensive rotational mechanics. A critical avenue for future work is extending this framework to 3D marker-less analysis. Integrating monocular 3D human mesh recovery models or lifting 2D representations into multi-view 3D coordinate space would eliminate perspective distortion. This advancement will ensure highly precise volumetric joint kinematics and afford a more comprehensive, biomechanically complete injury risk assessment tool.

6 Conclusion

We built a complete marker-less tennis serve analysis pipeline that segments recreational serve videos into eight biomechanical stages and uses those stage labels to drive interpretable injury-risk and technique checks. The best TAS configuration, MS-TCN++ with explicit pose/racket/ball features, ASRF-style boundary supervision, and constrained decoding, produced strong test-set stage segmentation and enabled a rule engine that generates visually grounded diagnostic reports. Although the current system remains limited by dataset size, single-view geometry, and racket/ball tracking reliability, it demonstrates a practical path toward accessible, reproducible, frame-level tennis biomechanics from ordinary video.

A Biomechanical Rule Set

Stage	Rule	Criterion
Start	1.1 Foot Width	Ankle width divided by shoulder width must be at least 0.80.
Start	1.2 Shoulder Level	Absolute shoulder tilt must be greater than 15 degrees.
Release	2.1 Elbow Rising	Serving elbow must be above the serving-side hip.
Release	2.2 Toss Arm Rising	Tossing wrist must be above the tossing shoulder.
Loading	3.1 Elbow Height	Serving elbow must be at or above serving shoulder level.
Loading	3.2 Elbow Angle	Serving elbow angle must be between 80 and 110 degrees.
Loading	3.3 Toss Arm Extension	Tossing elbow angle must be at least 150 degrees.
Loading	3.4 Knee Bend	Serving-side knee angle must be at most 160 degrees.
Cocking	4.1 Racket Drop	Serving wrist must be below the serving elbow.
Acceleration	5.1 Elbow Leading	Serving elbow must be above shoulder level at the start of acceleration.
Acceleration	5.2 Wrist Acceleration	Serving wrist must move above the elbow by the end of acceleration.
Acceleration	5.3 Leg Drive	Serving-side knee angle must be at least 155 degrees by the upward drive.
Contact	6.1 Arm Extension	Serving elbow angle at contact must be at least 160 degrees.
Contact	6.2 Contact Height	Contact proxy, preferably the cleaned ball annotation, must be above shoulder level.
Contact	6.3 Shoulder Rotation	Shoulder line must be rotated forward relative to the hip line at contact.
Deceleration	7.1 Cross-Body Follow-Through	Serving wrist must cross the torso midline during follow-through.
Deceleration	7.2 Elbow Deceleration	Serving elbow angle must flex to at most 130 degrees during follow-through.
Finish	8.1 Landing Knee Flexion (Serving Side)	Serving-side knee angle must be at most 160 degrees.
Finish	8.2 Landing Knee Flexion (Non-Dominant Side)	Non-dominant knee angle must be at most 165 degrees.

B Example Report for Test Serve 5

The following is the diagnostic report produced by the rule engine for test serve 5 using model-predicted stage labels, cached pose, and the cleaned manual racket/ball annotations.

```
Pose Analysis Report
Serves analyzed: 1

============================================================
SERVE 5  |  Result: In
============================================================
Sequence: test_serve_005
Stage labels: model predictions
Feature source: tas/manual_feature_annotations/test_serve_005.json
Passed: 11  |  Warnings: 6  |  Risks: 2

Predicted stage segments:
  start         frames   0-0
  release       frames   1-12
  loading       frames  13-20
  cocking       frames  21-34
  acceleration  frames  35-41
  contact       frames  42-42
  deceleration  frames  43-55
  finish        frames  56-61

=== Stage 1: Start ===
  PASS [1.1 Foot Width] Ankle width 2.06x shoulder width (meets >=0.80)
  PASS [1.2 Shoulder Level] Shoulder tilt 38.5 deg (meets >15 deg)

=== Stage 2: Release ===
  PASS [2.1 Elbow Rising] Serving elbow is 88px above hip;
```

```
      wind-up initiated
  WARN [2.2 Toss Arm Rising] Tossing wrist is 134px below tossing shoulder

=== Stage 3: Loading ===
  RISK [3.1 Elbow Height] Serving elbow is 44px below shoulder
  PASS [3.2 Elbow Angle] Serving elbow angle 83.1 deg in target range
  PASS [3.3 Toss Arm Extension] Toss arm extended at 173.5 deg
  WARN [3.4 Knee Bend] Serving knee too straight at loading (171.2 deg)

=== Stage 4: Cocking ===
  WARN [4.1 Racket Drop] Serving wrist is not below elbow

=== Stage 5: Acceleration ===
  PASS [5.1 Elbow Leading] Serving elbow is above shoulder
  PASS [5.2 Wrist Acceleration] Serving wrist moved 52px above elbow
  WARN [5.3 Leg Drive] Serving knee remains flexed at 137.8 deg

=== Stage 6: Contact ===
  PASS [6.1 Arm Extension] Serving arm extended at contact (162.2 deg)
  PASS [6.2 Contact Height] Contact proxy (ball) is 263px above
       shoulder level
  PASS [6.3 Shoulder Rotation] Shoulder line rotated 116.6 deg
       relative to hips

=== Stage 7: Deceleration ===
  RISK [7.1 Cross-Body Follow-Through] Racket-arm wrist does not
       cross midline
  WARN [7.2 Elbow Deceleration] Serving elbow still extended (166.5 deg)

=== Stage 8: Finish ===
  WARN [8.1 Landing Knee Flexion (Serving Side)] Serving knee too
       straight
  PASS [8.2 Landing Knee Flexion (Non-Dominant Side)]
       Non-dominant knee 160.2 deg
```

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Linfeng Dong, Yuchen Yang, Hao Wu, Wei Wang, Yuenan Hou, Zhihang Zhong, and Xiao Sun. Racketvision: A multiple racket sports benchmark for unified ball and racket analysis. *arXiv preprint arXiv:2511.17045*, 2026.

[3] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. *arXiv preprint arXiv:2007.06866*, 2020.

[4] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. Version 11.0.0.

[5] Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: Unified real-time end-to-end vision models, 2026.

[6] Mark Kovacs and Todd Ellenbecker. An 8-stage model for evaluating the tennis serve: Implications for performance enhancement and injury prevention. *Sports Health*, 3(6):504–513, 2011.

[7] Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *arXiv preprint arXiv:2006.09220*, 2020.

[8] Zijia Lu and Ehsan Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18175–18185, 2024.

[9] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[10] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *British Machine Vision Conference*, 2021.