TENNISRAM: ADVANTAGE-AWARE TENNIS COMMENTARY THROUGH JOINT PERCEPTION AND STATE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in multimodal large language models (MLLMs) have extended tennis video understanding from event recognition to match analysis and commentary generation. However, high-speed ball motion, transient racket–ball contact, and continuous player interactions pose challenges for fine-grained perception and temporal understanding. Existing methods typically rely on specialized models to detect ball locations, player positions, and stroke events, and convert these predictions into structured representations for commentary generation. Although such representations provide explicit event cues, they often overlook continuous visual evidence, such as stroke preparation, body balance, and recovery dynamics. To address these limitations, we propose **TennisRAM**. We first adopt high-frame-rate video sampling to preserve rapid motion information. Then, predictions from auxiliary perception heads for ball localization, player pose estimation, and stroke recognition are used to guide spatial feature compression, improving computational efficiency while enhancing fine-grained player–ball interaction representations. To model within-rally dynamics, we further introduce a **relative advantage state estimation** task, where the model determines which player holds the advantage and provides supporting evidence using only the video observed up to the current time step. The estimated advantage states and their rationales are incorporated with visual features as conditions for commentary generation. For training and evaluation, we annotate 1,200 tennis rallies with relative advantage states and point-formation labels, yielding 4,322 advantage-state annotations. TennisRAM achieves 66.33% accuracy on relative advantage state estimation, with ROUGE-L and CIDEr scores of 33.11 and 44.15, respectively, for commentary generation. Ablation studies demonstrate that fine-grained perceptual guidance and conditioning on both advantage evolution and point outcomes improve commentary quality.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have extended sports video understanding beyond predefined event recognition toward higher-level tasks such as match commentary and tactical analysis (Mkhallati et al., 2023; Rao et al., 2024; Liu et al., 2026). For tennis commentary, understanding the players' relative advantage is essential to explaining how a rally develops, rather than merely recognizing individual strokes. This requires fine-grained, temporally continuous visual evidence of the ball, players, and their interactions. However, video MLLMs typically process only a limited number of frames or visual tokens because of input-length and computational constraints (Xu et al., 2024; Islam et al., 2025). This limitation is particularly severe for tennis, where high-speed ball motion, transient racket–ball contact, and rapid player interactions require fine-grained spatiotemporal perception (Huang et al., 2019; Liu et al., 2025). Sparse temporal sampling may miss critical contact moments and abrupt trajectory changes. Meanwhile, highly compressed visual representations from the final layer of MLLM vision encoders may obscure spatial details of small objects such as the ball. These limitations weaken the visual evidence available for assessing relative advantage.

To compensate for these perceptual limitations, existing tennis understanding systems often adopt a perception–parsing–reasoning pipeline. External specialized models, such as TrackNet-based ball

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.

trackers, YOLO-based player detectors, and stroke recognizers, extract ball trajectories, player locations, and stroke events. These predictions are then converted into structured representations for downstream analysis and commentary generation (Liu et al., 2026). Although these representations provide explicit event cues, they may discard continuous visual evidence of stroke preparation, body balance, recovery dynamics, and open-court space. For instance, the same forehand may impose pressure when executed from a balanced position, yet represent a forced response when performed off balance. Thus, event sequences may describe *what* happened without fully explaining how each interaction changes the players' *relative advantage*. We characterize this relation at each moment as the *relative advantage state*, indicating which player, if either, currently holds the initiative, and associate it with a *supporting rationale* grounded in visual evidence. Tracking these states captures *relative advantage evolution*: how the advantage is gained, maintained, neutralized, or transferred. This connects local player–ball interactions to rally progression, providing process-level evidence for commentary generation.

To bridge fine-grained perception and process-level commentary, we propose **TennisRAM** (**R**elative **A**dvantage **M**odeling), a framework for tennis commentary. We introduce auxiliary supervision from ball localization, player pose estimation, and stroke recognition to learn fine-grained player–ball representations without relying on external perceptual preprocessing at inference. Ball and pose cues further support *perception-guided visual compression*, retaining task-relevant evidence within a limited token budget. Building on the compressed visual representations, the model predicts relative advantage states and supporting rationales from video prefixes, using only observations available up to each prediction timestamp. To connect relative advantage evolution to the final point outcome, we further predict *point-formation information* at rally completion, identifying the factors contributing to that outcome. Finally, compressed visual representations, relative advantage evolution, and point-formation information jointly condition commentary generation. To support training and evaluation, we augment TennisVL (Liu et al., 2026) with perceptual and analytical annotations. We automatically generate frame-level ball-position and player-pose annotations for all 43,228 videos, and manually annotate relative advantage states, supporting rationales, and point-formation information for 1,200 rallies from 10 matches.

2 Related work

Understanding fast rally sports. Recent sports video understanding has moved beyond predefined event recognition toward match description and analytical commentary (Mkhallati et al., 2023; Rao et al., 2024; Wang et al., 2026). For high-speed rally sports such as tennis, accurate understanding requires fine-grained perception of fast-moving small objects, transient racket–ball contact, and continuous player interactions. Prior work establishes such perceptual foundations through ball tracking, fine-grained event detection, and shot-level annotation (Huang et al., 2019; Liu et al., 2025; Wang et al., 2023), and further organizes perceptual outputs into structured representations for commentary generation (Liu et al., 2026). Modeling cross-shot relations, evidence chains, and action quality further improves the understanding of event dependencies and tactical implications (Mei et al., 2026; He et al., 2025). However, recognizing events and their relations alone is insufficient to characterize how a rally evolves. A model must also determine how each interaction changes the players' relative advantage, e.g., whether a return sustains pressure or relieves a disadvantaged situation. Such judgments require fine-grained visual evidence, including stroke preparation, body balance, and recovery dynamics, beyond action categories and spatial positions alone.

Perception-enhanced multimodal representations. Approaches for acquiring such fine-grained evidence can be broadly grouped into information injection and representation supervision. Information-injection methods use specialized models to provide player identities, spatial locations, or region-level features. These signals are then fused with global visual representations to strengthen entity–action associations (Xi et al., 2025; Li et al., 2026). Their effectiveness, however, depends on the intermediate representation and fusion strategy, while predefined event attributes may not preserve the continuous details of action execution and response. Representation-supervision methods instead improve spatial perception through localization objectives, region–language alignment, and local visual encoding (Chen et al., 2023; You et al., 2024; Ma et al., 2024). Geometry-enhanced approaches further inject spatial knowledge into internal representations through teacher-feature alignment or supervision from depth and camera geometry, reducing the need for external spatial teachers or additional geometric inputs at inference time (Li et al., 2025; Wu et al., 2026). Fol-

lowing this direction, we enhance perception during training using auxiliary supervision from ball and player localization, ball motion, and stroke recognition. This encourages fine-grained representations that can be directly used for match understanding while reducing reliance on external perceptual preprocessing at inference time.

Sports state and advantage modeling. Using fine-grained visual evidence for commentary further requires modeling how continuous interactions change the players' relative advantage. Related sports research mainly studies relational metrics and conditional value estimation. The former characterizes spatial advantage and temporal pressure from relative positioning and shot timing (van Meurs et al., 2021; Mlakar & Kovalchik, 2020), while the latter estimates scoring opportunities and their changes from event histories or motion trajectories (Cervone et al., 2014; Decroos et al., 2019; Kovalchik et al., 2020). More general video-understanding work studies process changes through joint action–state learning and state-tracking benchmarks (Souček et al., 2022; Yu et al., 2026). These studies provide useful foundations for process modeling, but quantifying positional relations or scoring opportunities does not directly yield semantic explanations suitable for commentary. We instead perform relative advantage state estimation on video prefixes and combine it with rally-level supervision on point-formation patterns. This allows the model to learn how local interactions change the players' relative advantage while explicitly separating within-rally advantage from the final point outcome. This formulation uses advantage evolution to connect local interactions, pressure changes, and defensive responses, providing a process-oriented structure for commentary generation.

3 DATASET

We construct the annotations through a combination of automatic and manual labeling. For fine-grained perception, frame-level ball positions and player poses are automatically generated for all 43,228 videos. For higher-level match understanding, we further select 1,200 rally videos covering 10 matches and manually annotate their relative advantage states and point-formation information. All videos are drawn from TennisVL (Liu et al., 2026).

Auxiliary perception annotations. Ball positions are obtained using WASB (Tarashima et al., 2023), with each frame represented by $b_t = (x_t, y_t, c_t, v_t)$, where (x_t, y_t) denotes the normalized ball position, c_t the tracking confidence, and v_t its visibility. Player poses are extracted using YOLOv8-Nano (Ultralytics, 2023) and RTMPose-M (Jiang et al., 2023) following the COCO 17-keypoint definition. The j-th keypoint at frame t is represented as $p_{t,j} = (x_{t,j}, y_{t,j}, s_{t,j})$, where $(x_{t,j}, y_{t,j})$ denotes the normalized image coordinates and $s_{t,j}$ the keypoint confidence. Stroke labels are inherited from TennisVL and temporally aligned with the corresponding frames.

Point-formation annotations. For each of the 1,200 manually annotated rallies, we label the point winner and the causes of point formation. The winner is categorized as the near- or far-side player, while the causes are represented by a multi-label set of 12 categories covering serve and return, attacking patterns, stroke properties, and error types. To avoid temporal and player overlap across subsets, we split the data by match identity, using six matches for training (727 rallies), two for validation (254 rallies), and two for testing (219 rallies). The full label taxonomy is provided in the supplementary material.

Relative advantage state annotations. To further capture how the rally evolves before the final outcome, we annotate relative advantage states at approximately 2-second intervals. Each timestamp is labeled as near-side advantage, far-side advantage, or balanced. For non-balanced states, annotators additionally select supporting rationales from 5 advantage and 5 disadvantage factors covering stroke preparation, court space, positional control, and offensive–defensive initiative. Two tennis experts independently annotate both the point-formation and relative-advantage labels, and disagreements are reviewed and resolved to obtain the final annotations. Detailed label definitions are provided in the supplementary material.

4 PROBLEM DEFINITION

Given a tennis rally video $\mathcal{V}$, we formulate relative advantage modeling as the temporal prediction of the within-rally relative advantage state and its supporting rationale. Let $\mathcal{V}_{\leq \tau_t}$ denote the video

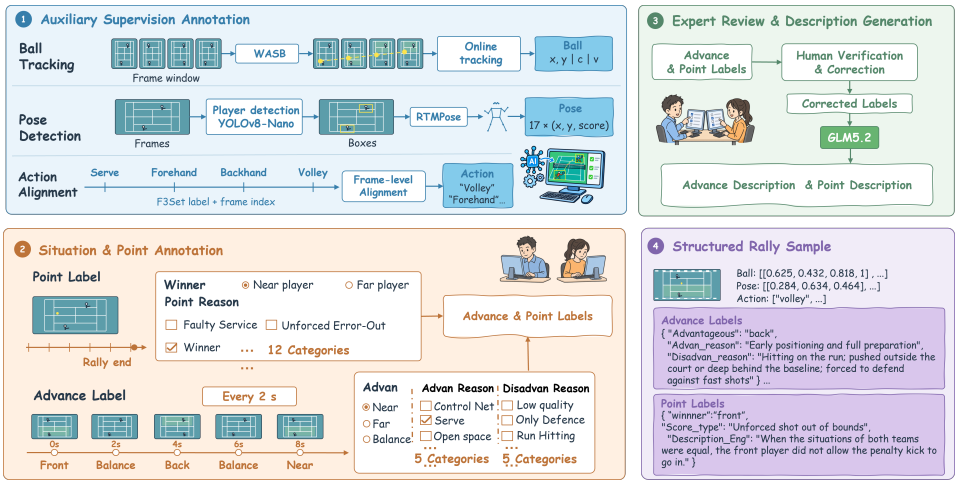

Figure 1: Dataset construction pipeline.

prefix observed up to time τ_t. The relative advantage state at τ_t is defined as s_t^a and s_t^r, where $s_t^a \in \{\text{near}, \text{neutral}, \text{far}\}$ indicates near-side advantage, a neutral state, or far-side advantage, and s_t^r is a multi-label rationale vector describing the factors underlying the current advantage state.

For a complete rally, let w denote the point winner and c is the multi-label point-formation vector. Each dimension corresponds to a point-formation cause or tactical attribute, with multiple labels allowed simultaneously. The two prediction tasks are formulated as

$$(\hat{s}_t^a, \hat{s}_t^r) = f_\theta(\mathcal{V}_{\leq \tau_t}), \qquad \hat{c} = g_\theta(w, \mathcal{V}_{\leq \tau_\star}), \tag{1}$$

where $\tau_\star$ denotes the end of the rally. Thus, $(\hat{a}_t, \hat{r}_t)$ describes the evolving relative advantage within the rally, while $(w, \hat{c})$ summarizes the final point outcome and its formation causes.

5 Method

We build **TennisRAM** on InternVL3.5-8B (Wang et al., 2025) with auxiliary perception, perception-guided token compression, temporal advantage and point-formation modeling, and commentary generation. Fine-grained ball, pose, and stroke cues from the high-frame-rate branch guide visual compression, while the resulting compact features support temporal state prediction and, together with the predicted analytical information, condition final commentary generation Figure 2.

5.1 Auxiliary Perception Heads

To enhance fine-grained player–ball representations, we introduce auxiliary supervision from ball localization, player pose estimation, and stroke recognition. These tasks share a common perception feature while using task-specific prediction heads.

For frame t, let $F_t^{(\ell)} \in \mathbb{R}^{d_v \times H \times W}$ denote the patch features from layer ℓ of the visual encoder. We reshape features from the 12th-, 6th-, and 1st-to-last layers to the spatial patch grid, align their channels using 1×1 convolutions, and sum them. The fused feature is further refined by two 3×3 convolutional blocks to obtain the shared perception feature $U_t \in \mathbb{R}^{256 \times H \times W}$.

Based on U_t, the ball localization and pose estimation heads use separate 1×1 convolutional readouts to predict a ball heatmap $\hat{B}_t \in \mathbb{R}^{H \times W}$ and player pose heatmaps $\hat{P}_t \in \mathbb{R}^{2 \times 17 \times H \times W}$, respectively.

To incorporate player posture into stroke recognition, we first aggregate the keypoint heatmaps of each player $p \in \{\text{near}, \text{far}\}$ into a spatial attention map:

$$A_{t,p} = \text{softmax}_{HW}\left(\beta \max_j \hat{P}_{t,p,j}\right), \tag{2}$$

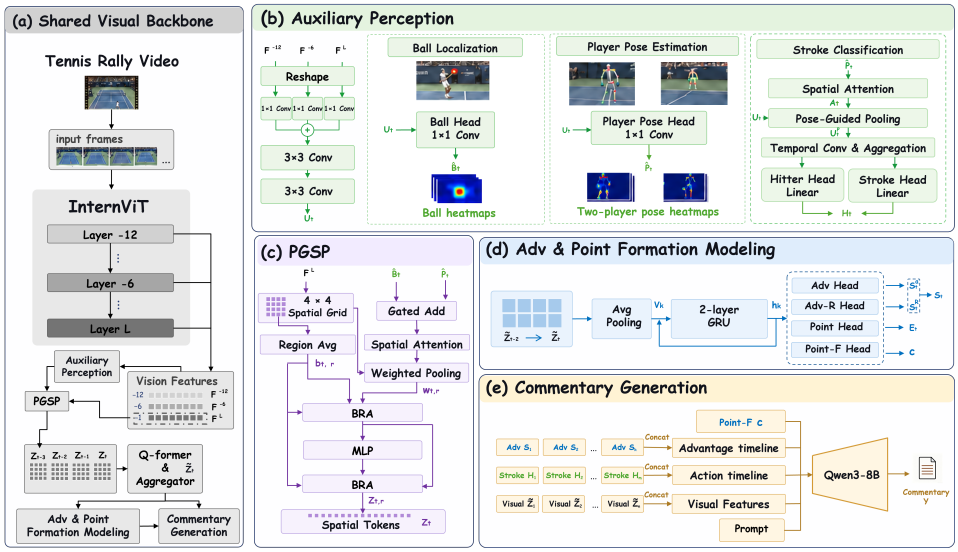

Figure 2: Overview of TennisRAM. (a) Multi-level visual features are extracted from rally videos. (b) Auxiliary perception learns fine-grained ball, pose, and stroke representations. (c) Visual tokens are compressed with perception guidance. (d) Rally advantage and point formation are modeled recurrently. (e) Conditioned on visual and analytical signals, the model generates tennis commentary.

where $\hat{P}_{t,p,j} \in \mathbb{R}^{H\times W}$ denotes the predicted heatmap of the j-th keypoint for player p at frame t. The max operation over keypoints aggregates individual responses into a player-level spatial map, β controls the sharpness of the attention distribution, and $\mathrm{softmax}_{HW}$ normalizes responses over the spatial grid. The resulting $A_{t,p} \in \mathbb{R}^{H\times W}$ therefore highlights regions associated with player p.

We then use this attention map to pool the shared perception feature:

$$U^P_{t,p} = \sum_{h=1}^{H}\sum_{w=1}^{W} A_{t,p}(h,w)\, U_t(:,h,w), \tag{3}$$

where $U_t(:,h,w) \in \mathbb{R}^{256}$ denotes the shared perception feature at spatial location (h,w). The resulting $U^P_{t,p} \in \mathbb{R}^{256}$ is a pose-guided representation of player p at frame t.

Finally, we combine $U^P_{t,p}$ with the global frame feature and a learnable player-side embedding, followed by a shared temporal encoder to capture player motion dynamics. The resulting player representation is used by two lightweight heads to predict the hitter identity and stroke category, whose outputs are combined for the final stroke prediction H_t.

5.2 Perception-Guided Spatiotemporal Token Compression

To preserve fine-grained motion information while reducing the visual sequence length, we apply Perception-Guided Spatial Pooling (PGSP) to sampled-frame visual features, followed by Q-Former-based temporal aggregation.

Given the final-layer visual feature $F^{(L)}_t \in \mathbb{R}^{D\times H\times W}$, where t denotes the frame index, L denotes the final layer of the visual encoder, and D, H, W represent the feature dimension and spatial resolution, PGSP partitions the spatial feature map into a 4×4 grid. The ball-position and player-pose heatmaps are resized to the feature resolution and normalized separately. The normalized heatmaps are weighted by learnable gates and combined to obtain the spatial response map A_t.

For the r-th spatial region $\mathcal{R}_r$, we first compute the regional mean feature:

$$b_{t,r} = \frac{1}{|\mathcal{R}_r|}\sum_{q\in\mathcal{R}_r} F^{(L)}_{t,q}, \tag{4}$$

where r denotes the region index and q denotes the spatial location.

The response map is normalized within each region to generate spatial attention weights $\alpha_{t,q}$, which are used to compute the perception-weighted feature $w_{t,r}$:

$$w_{t,r} = \sum_{q \in \mathcal{R}_r} \alpha_{t,q} F_{t,q}^{(L)}, \qquad \alpha_{t,q} = \frac{\exp(A_{t,q})}{\sum_{s \in \mathcal{R}_r} \exp(A_{t,s})}, \tag{5}$$

where $A_{t,q}$ denotes the response value at spatial location q.

Using the regional mean feature as a reference, PGSP injects perceptual information through bounded residual adaptation:

$$\begin{aligned} l_{t,r} &= b_{t,r} + \mathcal{B}_{b_{t,r}}(w_{t,r} - b_{t,r}), \\ z_{t,r} &= b_{t,r} + \mathcal{B}_{b_{t,r}}(\mathrm{MLP}(l_{t,r})), \end{aligned} \tag{6}$$

where $\mathcal{B}_b(\cdot)$ constrains the RMS magnitude of the residual to 10% of the reference feature magnitude.

After PGSP, each frame feature Z is represented by 16 spatially structured tokens: We then group every four consecutive frames into a non-overlapping temporal window. A Q-Former with M learnable queries aggregates the $64 \times D$ tokens within each window into $M \times D$ temporal tokens. The window-level representations $\widetilde{Z}$ are concatenated chronologically, which serves as the visual input for relative advantage modeling and commentary generation.

5.3 Temporal Relative Advantage Modeling

To model the evolution of relative advantage within a rally, we perform temporal modeling over spatiotemporally compressed visual features using a 2-second observation window and a 1-second prediction interval.

The compressed visual sequence contains four window-level representations per second, each consisting of M visual tokens with dimension D. Starting from the 2-second timestamp, we make predictions at each timestamp τ_k and collect all tokens within $(\tau_k - 2, \tau_k]$:

$$V_k \in \mathbb{R}^{8M \times D}. \tag{7}$$

The tokens within each observation window are average-pooled and projected into a 256-dimensional representation e_k. The resulting representations are sequentially processed by a two-layer unidirectional GRU:

$$h_k = \mathrm{GRU}(e_k, h_{k-1}), \qquad h_k \in \mathbb{R}^{256}, \tag{8}$$

where the initial hidden state is initialized to zero. The recurrent state captures historical rally context while each prediction incorporates the latest 2-second visual evidence.

Based on the shared hidden state h_k, task-specific heads predict the relative advantage state, advantage rationale, rally termination, and point-formation outcomes. For a complete rally, the terminal point-winner prediction determines the final result, while temporal advantage predictions are filtered by confidence and converted into timestamped textual conditions for commentary generation.

5.4 Commentary Generation

The commentary branch generates natural-language commentary by combining compressed visual features with structured match information. Specifically, the compressed visual tokens $\widetilde{Z}$ are concatenated with a prompt containing timestamped relative advantage states, point-formation information, and stroke actions. The resulting multimodal input is encoded by the LLM to generate the final commentary Y.

5.5 Training Strategy

We train TennisRAM in three stages.

Stage I: Auxiliary perception pretraining. We freeze the InternVL backbone and train only the auxiliary perception heads. The objective jointly supervises ball localization, player pose estimation, stroke recognition, and hitter classification:

$$\mathcal{L}_{\text{aux}} = \lambda_{\text{b}}\mathcal{L}_{\text{b}} + \lambda_{\text{p}}\mathcal{L}_{\text{p}} + \lambda_{\text{act}}\mathcal{L}_{\text{action}} + \lambda_{\text{h}}\mathcal{L}_{\text{hitter}}, \tag{9}$$

where $\mathcal{L}_{\text{b}}$ and $\mathcal{L}_{\text{p}}$ are weighted mean-squared errors for ball and pose heatmaps, and $\mathcal{L}_{\text{action}}$ and $\mathcal{L}_{\text{hitter}}$ are cross-entropy losses for stroke and hitter classification. We set $\lambda_{\text{b}} = \lambda_{\text{p}} = 1$ throughout training. For epochs 1–20, $\lambda_{\text{act}} = \lambda_{\text{h}} = 0$; for epochs 21–40, we set $\lambda_{\text{act}} = 2$ and $\lambda_{\text{h}} = 1$.

Stage II: Relative advantage modeling. We freeze the InternVL backbone and auxiliary perception heads, and train the visual compression and projection modules together with the relative advantage prediction heads. The objective is

$$\mathcal{L}_{\text{state}} = \mathcal{L}_{\text{adv}} + \lambda_{\text{advr}}\mathcal{L}_{\text{advr}} + \lambda_{\text{ed}}\mathcal{L}_{\text{ed}} + \lambda_{\text{pf}}\mathcal{L}_{\text{pf}}, \tag{10}$$

where $\mathcal{L}_{\text{adv}}$ is the cross-entropy loss for three-way relative advantage classification, $\mathcal{L}_{\text{advr}}$ is the binary cross-entropy loss for multi-label advantage rationales, $\mathcal{L}_{\text{ed}}$ is the binary cross-entropy loss for rally termination prediction, and $\mathcal{L}_{\text{pf}}$ is the binary cross-entropy loss for multi-label point-formation causes. We set $\lambda_{\text{advr}} = 1.0$, $\lambda_{\text{ed}} = 0.5$, and $\lambda_{\text{pf}} = 1.0$. This stage is trained for 30 epochs.

Stage III: Commentary adaptation. Starting from the Stage-II model, we jointly optimize relative advantage modeling and commentary generation, while training the rank-32 LoRA adapters of the language decoder. The commentary objective is

$$\mathcal{L}_{\text{cap}} = -\frac{1}{|Y|}\sum_{i=1}^{|Y|}\log p(y_i \mid \mathcal{X}, y_{<i}), \tag{11}$$

where Y denotes the target commentary, y_i is its i-th token, and $\mathcal{X}$ denotes the multimodal input to the language model. The overall objective is

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{state}} + \lambda_{\text{cap}}\mathcal{L}_{\text{cap}}, \tag{12}$$

where λ_{cap}=3.0.

6 Experiments

6.1 Experimental setup

Tasks and implementation. TennisRAM is initialized from the pretrained weights of InternVL3.5-8B. In Stage I, we pretrain the auxiliary perception modules for 40 epochs using AdamW with a batch size of 16 and a learning rate of 2×10^{-4}. In Stage II, the relative advantage head is trained for 30 epochs with a learning rate of 1×10^{-4} and the same batch size. In Stage III, we adapt the language model for 2 epochs using rank-32 LoRA with a learning rate of 1×10^{-5} and a batch size of 16. All experiments are conducted on 4 × NVIDIA H800 80GB GPUs.

Evaluation Metrics. We adopt task-specific metrics to evaluate the different prediction outputs. Advantage prediction (Adv.) is evaluated with accuracy, while multi-label advantage rationales (Adv.-R) and point-formation causes (Point-R) are evaluated with accuracy and recall. Commentary generation is evaluated using BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015).

Baselines. We compare against the general-purpose MLLMs Qwen3-VL-8B (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025), GPT-5.2 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), and Gemini 3 Flash (Doshi, 2025), all evaluated in zero-shot settings, together with the tennis-specific TennisExpert (Liu et al., 2026). For commentary generation, we reuse the results reported by (Liu et al., 2026), while all other results are produced through our own experiments.

6.2 Comparison with Existing Methods

We compare TennisRAM with existing methods on relative advantage estimation, point-formation causes, and commentary generation. For the advantage-related tasks, the TennisExpert baseline

Table 1: State, outcome, and commentary results. Classification metrics are percentages; bold indicates the largest value in each column. Commentary baselines marked † match the published TennisExpert results (Liu et al., 2026). TennisExpert state results use the adaptation described in the text.

	Adv.	Adv.-R		Point-R		Commentary		
Method	Acc.	Acc.	Recall	Acc.	Recall	BLEU-4	ROUGE-L	CIDEr
GPT-5.2†	40.21	44.55	47.73	41.71	39.21	1.32	15.95	7.52
Claude Sonnet 4.5†	38.14	37.12	29.90	42.19	31.14	1.22	14.20	2.09
Gemini 3 Flash†	27.10	29.23	31.19	33.06	28.01	1.57	16.02	7.31
Qwen3-VL-8B†	11.79	19.80	15.61	21.12	20.13	1.20	14.90	3.89
InternVL3.5-8B†	13.12	17.02	20.21	22.11	19.17	0.75	13.61	0.08
TennisExpert†	41.78	47.95	45.31	51.93	45.20	**7.98**	29.16	43.71
TennisRAM	**61.83**	**66.33**	**62.90**	**74.67**	**71.63**	6.51	**33.11**	**44.15**

combines event text available up to each prediction timestamp with visual features sampled at 2 FPS, following its original setting, and trains the same advantage-head architecture used in our framework. All general-purpose MLLMs are evaluated zero-shot.

Relative advantage and point-formation prediction. As shown in Table 1, TennisRAM achieves 61.83% accuracy on advantage-state prediction, outperforming the adapted TennisExpert baseline by 20.05 percentage points (61.83% vs. 41.78%). For advantage rationales, TennisRAM obtains 66.33% accuracy and 62.90% recall, improving by 18.38 and 17.59 points, respectively. For point-formation causes, it achieves 74.67% accuracy and 71.63% recall, with gains of 22.74 and 26.43 points. The consistent improvements in both advantage-state and rationale prediction suggest that TennisRAM better captures not only which player holds the advantage, but also the evidence underlying that state.

Commentary generation. TennisRAM achieves ROUGE-L and CIDEr scores of 33.11 and 44.15, exceeding the values reported for TennisExpert by 3.95 and 0.44 points, respectively, while its BLEU-4 score is slightly lower (6.51 vs. 7.98). The improvement does not hold uniformly across all generation metrics. Since the TennisExpert commentary scores are taken from its original paper, these comparisons should be interpreted subject to differences in evaluation settings and pipelines.

Among the zero-shot general-purpose MLLMs, GPT-5.2 obtains the highest advantage-state accuracy of 40.21%, while TennisRAM reaches 61.83%. This gap indicates the benefit of task-specific training for relative advantage estimation. We further examine the contributions of individual components through ablation studies in the following sections.

6.3 Ablation on Auxiliary Perception

To examine how different perceptual cues contribute to relative advantage estimation and commentary generation, we ablate the Ball, Pose, and Action branches under six configurations, with the results reported in Table 2. When a branch is disabled, we remove its supervision loss, prediction module, and downstream inputs. Since action recognition relies on pose-guided player representations, all configurations with Action also retain Pose.

All auxiliary configurations outperform the baseline. Using Pose alone performs better than Ball alone on all five relative-advantage and point-formation metrics; in particular, the recall of advantage rationales and point-formation causes improves by 8.71 and 7.69 percentage points, respectively. This suggests that player pose provides useful cues about positioning and preparation for relative advantage estimation. By contrast, Ball alone is more effective for commentary generation, improving CIDEr by 7.28 points over Pose alone. This result suggests that ball tracking plays an important role in commentary generation. Combining Ball and Pose further improves metrics over either single branch, suggesting that ball dynamics and player states provide complementary cues for relative advantage estimation and commentary generation.

Table 2: Auxiliary perception ablation. Checkmarks indicate enabled supervision, prediction modules, and associated inputs. Bold marks column maxima.

Auxiliary signals			Adv.	Adv.-R		Score-R		Commentary		
Ball	Pose	Action	Acc.	Acc.	Recall	Acc.	Recall	BLEU-4	ROUGE-L	CIDEr
–	–	–	34.14	20.09	32.02	38.22	48.09	1.92	17.80	5.49
✓	–	–	50.12	37.11	44.21	45.50	57.95	5.60	28.35	32.45
✓	✓	–	**63.12**	60.90	**64.27**	65.42	64.21	6.12	30.20	42.81
–	✓	–	55.17	50.82	57.02	53.19	66.11	5.24	19.12	25.17
–	✓	✓	57.84	53.19	53.56	58.45	56.66	5.43	24.91	29.78
✓	✓	✓	61.83	**66.33**	62.90	**74.67**	**71.63**	**6.51**	**33.11**	**44.15**

Table 3: Commentary text-condition ablation. Adv.(-R) includes advantage and its supporting attributes; Point(-R) includes the point winner and predicted outcome attributes.

Adv.(-R)	Point(-R)	BLEU-4	ROUGE-L	CIDEr
–	–	5.98	25.09	37.91
✓	–	5.73	25.39	37.01
–	✓	**6.94**	28.05	41.35
✓	✓	6.51	**33.11**	**44.15**

6.4 Ablation on Advantage and Score Conditions

To examine whether relative advantage and point-formation information benefit commentary generation, we ablate two textual conditions: Adv.(-R), which contains the predicted advantage state and its rationale, and Score, which describes the final point outcome. We evaluate each condition individually and compare them with their combination.

Both conditions improve over the baseline, suggesting that process-level advantage information and final outcome information are useful for commentary generation. Using Score alone achieves BLEU-4, ROUGE-L, and CIDEr scores of 6.94, 28.05, and 41.35, respectively, outperforming the Adv.(-R)-only setting on all three metrics. This suggests that, when used alone, final point information aligns more directly with the reference commentary.

Combining Adv.(-R) and Score further improves ROUGE-L and CIDEr to 33.11 and 44.15, corresponding to gains of 5.06 and 2.80 over Score alone. Although BLEU-4 decreases slightly, this may reflect changes in the wording or organization of the generated commentary after incorporating within-rally advantage information, leading to lower local n-gram overlap with the reference commentary. Meanwhile, the improvements in ROUGE-L and CIDEr suggest that advantage states and their rationales provide complementary process-level information even when the final outcome is already available, consistent with our goal of linking rally evolution to point formation.

7 Conclusion

We present **TennisRAM**, an advantage-aware framework that connects fine-grained player–ball perception with within-rally state modeling for tennis commentary generation. TennisRAM uses auxiliary perception supervision and perception-guided compression to support relative advantage and point-formation modeling, whose outputs are incorporated into commentary generation. Experiments show improved advantage and rationale prediction and higher ROUGE-L and CIDEr scores, while ablations demonstrate complementary contributions from fine-grained perception and analytical conditions. Overall, these results show that within-rally advantage provides a useful intermediate representation for linking local interactions to process-aware tennis commentary.

AI USE STATEMENT

In this work, we used generative AI tools for literature retrieval and organization, writing assistance, debugging support, and synthetic text generation. Specifically, we used these tools to identify and organize related work on sports video understanding and to synthesize methodological trends, helping us formulate the research questions addressed in this work. During manuscript preparation, we used AI tools to check grammar and polish sentences. Three authors reviewed the AI-assisted revisions and confirmed their accuracy. During code development, we used AI tools only to analyze error messages and assist in diagnosing potential causes of bugs. During dataset construction, we used AI tools to generate descriptions of the reasons for tennis rally advantages and point outcomes in different writing styles. Human annotators reviewed these descriptions and confirmed their correctness. We take responsibility for the final content of this work, including the text, claims, code, and data produced or revised with the assistance of generative AI.

REFERENCES

Anthropic. Introducing Claude Sonnet 4.5. Official model announcement, 2025. URL https://www.anthropic.com/news/claude-sonnet-4-5. Accessed September 17, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry. A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes. *arXiv preprint arXiv:1408.0777*, 2014. URL https://arxiv.org/abs/1408.0777.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*, 2023. URL https://arxiv.org/abs/2306.15195.

Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions Speak Louder Than Goals: Valuing Player Actions in Soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019. URL https://www.janvanhaaren.be/assets/papers/kdd-2019-vaep.pdf.

Tulsee Doshi. Gemini 3 Flash: frontier intelligence built for speed. Google official model announcement, 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/. Accessed September 17, 2026.

Xusheng He, Wei Liu, Shanshan Ma, Qian Liu, Chenghao Ma, and Jianlong Wu. FineBadminton: A Multi-Level Dataset for Fine-Grained Badminton Video Understanding. *arXiv preprint arXiv:2508.07554*, 2025. URL https://arxiv.org/abs/2508.07554.

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. *arXiv preprint arXiv:1907.03698*, 2019. URL https://arxiv.org/abs/1907.03698.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. BIMBA: Selective-Scan Compression for Long-Range Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.09590.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399*, 2023. URL https://arxiv.org/abs/2303.07399.

Stephanie Kovalchik, Martin Ingram, Kokum Weeratunga, and Cagatay Goncu. Space-Time VON CRAMM: Evaluating Decision-Making in Tennis with Variational generatiON of Complete Resolution Arcs via Mixture Modeling. *arXiv preprint arXiv:2005.12853*, 2020. URL https://arxiv.org/abs/2005.12853.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial Forcing: Implicit Spatial Representation Alignment for Vision-language-action Model. *arXiv preprint arXiv:2510.12276*, 2025. URL https://arxiv.org/abs/2510.12276.

Yizhi Li, Jiawei Jiang, Guanhong Wang, Yingcai Wu, and Gaoang Wang. SportsGrounder: Proposal-Aided Interleaved Grounding for Dense Sports Video Reasoning. *arXiv preprint arXiv:2608.07932*, 2026. URL https://arxiv.org/abs/2608.07932.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 2004. URL https://aclanthology.org/W04-1013/.

Zhaoyu Liu, Kan Jiang, Murong Ma, Zhe Hou, Yun Lin, and Jin Song Dong. F^3Set: Towards Analyzing Fast, Frequent, and Fine-grained Events from Videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vlg5WRKHxh.

Zhaoyu Liu, Xi Weng, Lianyu Hu, Zhe Hou, Kan Jiang, Jin Song Dong, and Yang Liu. TennisExpert: Towards Expert-Level Analytical Sports Video Understanding. *arXiv preprint arXiv:2603.13397*, 2026. URL https://arxiv.org/abs/2603.13397.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized Visual Tokenization for Grounding Multimodal Large Language Models. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/01023.pdf.

Yifan Mei, Qingling Shi, Changli Wu, Jiayuan Rao, Jiayi Ji, and Liujuan Cao. TennisVAR: A Stroke-Evidence-Grounded Multimodal Large Language Model for Tactical Reasoning in Tennis Videos. *arXiv preprint arXiv:2608.12920*, 2026. URL https://arxiv.org/abs/2608.12920.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-Caption: Dense Video Captioning for Soccer Broadcasts Commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. URL https://arxiv.org/abs/2304.04565.

Miha Mlakar and Stephanie Ann Kovalchik. Analysing time pressure in professional tennis. *Journal of Sports Analytics*, 2020. doi: 10.3233/JSA-200406. URL https://journals.sagepub.com/doi/10.3233/JSA-200406.

OpenAI. Introducing GPT-5.2. Official model announcement, 2025. URL https://openai.com/index/introducing-gpt-5-2/. Accessed September 17, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. MatchTime: Towards Automatic Soccer Game Commentary Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.99/.

Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the Change: Learning Object States and State-Modifying Actions From Untrimmed Web Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.11637.

Shuhei Tarashima, Muhammad Abdul Haq, Yushan Wang, and Norio Tagawa. Widely Applicable Strong Baseline for Sports Ball Detection and Tracking. In *British Machine Vision Conference*, 2023. URL https://github.com/nttcom/WASB-SBDT.

Ultralytics. Ultralytics YOLOv8. Official model documentation, 2023. URL https://docs.ultralytics.com/models/yolov8. Accessed September 17, 2026.

Edda van Meurs, Tim Buszard, Stephanie Kovalchik, Damian Farrow, and Machar Reid. Interpersonal coordination in tennis: Assessing the Positional Advantage Index with Australian Open Hawkeye data. *International Journal of Performance Analysis in Sport*, 2021. doi: 10.1080/24748668.2020.1843213. URL https://cris.uni-muenster.de/portal/en/publication/77914300.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. URL https://arxiv.org/abs/1411.5726.

Kaiwen Wang, Kaili Zheng, Rongrong Deng, Yiming Shi, Chenyi Guo, and Ji Wu. BoxComm: Benchmarking Category-Aware Commentary Generation and Narration Rhythm in Boxing. *arXiv preprint arXiv:2604.04419*, 2026. URL https://arxiv.org/abs/2604.04419.

Wei-Yao Wang, Yung-Chang Huang, Tsi-Ui Ik, and Wen-Chih Peng. ShuttleSet: A Human-Annotated Stroke-Level Singles Dataset for Badminton Tactical Analysis. *arXiv preprint arXiv:2306.04948*, 2023. URL https://arxiv.org/abs/2306.04948.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Jing Wu, Jianhua Wu, Jiayi Guan, Jiahong Chen, Jinghui Lu, Hangjun Ye, Bingzhao Gao, and Long Chen. SpatioLM: Towards General Physical Spatial Intelligence in Vision-Language Models. *arXiv preprint arXiv:2608.01899*, 2026. URL https://arxiv.org/abs/2608.01899.

Zeyu Xi, Haoying Sun, Yaofei Wu, Junchi Yan, Haoran Zhang, Lifang Wu, Liang Wang, and Changwen Chen. Player-Centric Multimodal Prompt Generation for Large Language Model Based Identity-Aware Basketball Video Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2507.20163.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. SlowFast-LLaVA: A Strong Training-Free Baseline for Video Large Language Models. *arXiv preprint arXiv:2407.15841*, 2024. URL https://arxiv.org/abs/2407.15841.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/pdf/2334b28229d561c4b8239640b86f836796cdfa2f.pdf.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, and Saining Xie. Benchmarking Visual State Tracking in Multimodal Video Understanding. *arXiv preprint arXiv:2606.03920*, 2026. URL https://arxiv.org/abs/2606.03920.

.1 Annotation Categories for Rally Outcomes and Player States

Our annotation scheme describes completed rallies and intermediate player states at two temporal levels. Rally-level annotations capture point-ending events and relevant patterns of play. State-level annotations describe favorable and unfavorable conditions at the annotated timestamp. These categories provide interpretable attributes for commentary generation.

Rally-outcome categories. We define 12 categories covering errors, winning shots, and patterns associated with the point outcome:

1. **Service error:** A service fault associated with the point outcome. A first-service fault alone does not constitute a lost point; a double fault does.
2. **Winner:** A legal shot that wins the point without the opponent making contact with the ball.
3. **Change of shot direction:** Redirecting the ball relative to the preceding exchange, such as switching from cross-court to down-the-line play.
4. **Pace pressure:** Fast shots that reduce the opponent's available response time and constrain the return.
5. **Net attack:** Attacking play near the net, including volleys and overheads.
6. **Repeated placement:** Repeatedly directing shots toward the same court area to sustain pressure.
7. **Spin:** Salient use of spin that affects the ball's flight or bounce and the opponent's response.
8. **Unforced error:** A failed shot judged to occur without substantial pressure from the opponent's preceding shot.
9. **Ball hit out:** A shot that lands outside the valid court boundaries.
10. **Ball hit into the net:** A shot that fails to clear the net.
11. **Extended rally:** A sustained exchange involving multiple successive shots before the point ends.
12. **Depth pressure:** Deep shots landing near the opponent's baseline that restrict forward movement or comfortable contact.

Advantage-related categories. We define five categories describing favorable conditions for a player:

1. **Serve advantage:** A favorable position in the exchange established by the serve, such as a constrained return from the opponent.
2. **Control net:** Occupying a position near the net that enables attacking shots and limits the opponent's response options.
3. **Early positioning:** Reaching the anticipated hitting position with sufficient time to prepare for the incoming ball.
4. **Defensive counterattack:** Turning a defensive situation into an attacking opportunity through the return.
5. **Open space in the opponent's court:** A clearly exposed court area that provides a viable attacking target.

Disadvantage-related categories. We define five categories describing unfavorable conditions or a terminal outcome for a player:

1. **Run hitting:** Striking the ball while moving to reach it, with limited opportunity to establish a stable hitting position.
2. **Forced back to the baseline:** Being pushed into a deep defensive position at or behind the baseline, restricting opportunities to move forward and attack.

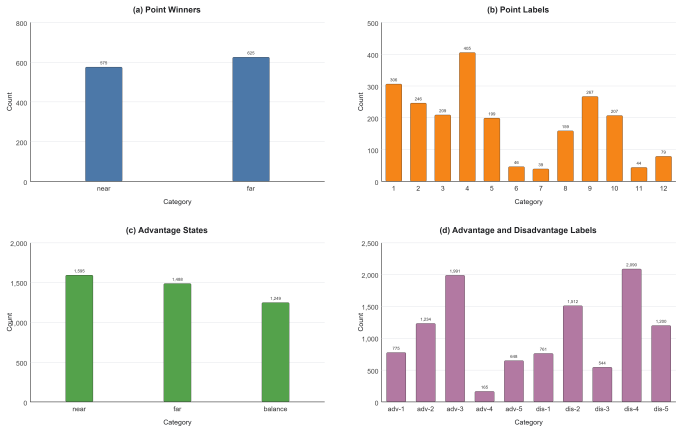

Figure 3: Label distributions of the annotated dataset. We report the numbers of annotations for the point winner, point-formation causes, relative advantage states, and advantage rationales.

3. **Low-quality return:** A return that provides insufficient depth, pace, or placement to limit the opponent's next attack.
4. **Only defence:** Being limited to a defensive response because the incoming ball leaves insufficient preparation time.
5. **Point lost:** The player has lost the completed point. This is a terminal label and does not describe a disadvantage before the rally ends.

.2 Additional Dataset Statistics

To provide a more detailed characterization of the annotated dataset, we further analyze the label distributions of the four annotation components used in our framework: the point winner, point-formation causes, relative advantage states, and advantage rationales. Specifically, we count the number of occurrences of each label over the annotated rallies and intermediate timestamps. These statistics complement the annotation definitions in the main paper by illustrating the empirical frequency of different point outcomes and within-rally advantage patterns in the dataset. The resulting distributions are shown in Fig. 3.

.3 Direct Heatmap Injection vs. Perception-Guided Compression

Our main method uses auxiliary perceptual signals to guide spatial feature compression, rather than directly providing the predicted heatmaps to the language model. To further examine the benefit of this design, we compare our perception-guided compression strategy with an alternative that directly injects the perceptual heatmaps into the LLM.

For the *Direct Heatmap Input* variant, we retain the same auxiliary ball localization, player pose estimation, and stroke recognition heads and optimize them using the same supervision as in our full model. However, their predicted heatmaps are not used to guide spatial feature compression. Instead, we replace the spatial guidance map used during compression with an all-one matrix, such that the compression module does not receive any spatial preference from the auxiliary perception predictions. Meanwhile, the predicted ball and pose heatmaps are projected into the language-model embedding space using an MLP and converted into additional visual tokens. These heatmap tokens are then concatenated with the original multimodal input and directly provided to the LLM. All other training settings and optimization strategies are kept identical to those of our proposed method, ensuring that the comparison isolates the effect of how auxiliary perceptual information is incorporated.

Table 4 reports the commentary-generation results. Directly injecting the heatmap representations obtains a BLEU-4 score of 5.18, a ROUGE-L score of 30.14, and a CIDEr score of 36.20. In

Table 4: Comparison between direct heatmap injection and perception-guided feature compression for commentary generation. All variants use the same auxiliary perception supervision and training strategy; they differ only in how the predicted perceptual information is incorporated into the visual representation.

Method	BLEU-4	ROUGE-L	CIDEr
Direct Heatmap Input	5.18	30.14	36.20
Perception-Guided Compression	**6.51**	**33.11**	**44.15**

comparison, our perception-guided feature compression achieves 6.51, 33.11, and 44.15 on the three metrics, respectively.

The consistent improvements across all three commentary metrics indicate that auxiliary perceptual information is more effective when it is used to enhance and selectively compress the underlying visual representation, rather than being appended to the LLM as additional projected tokens. In particular, perception-guided compression improves CIDEr from 36.20 to 44.15 and ROUGE-L from 30.14 to 33.11, suggesting that spatially incorporating fine-grained ball and player cues before compression better preserves information relevant to semantic and contextual commentary generation. These results provide additional evidence for our design choice of using auxiliary perception as representation guidance: instead of treating heatmaps as an independent input modality, our method exploits them to determine which spatial visual evidence should be retained during token compression.

.4 Qualitative examples

Figure 4 presents three rally examples with predicted relative advantage states and supporting rationales at sampled timestamps, together with point-formation predictions at rally completion. We compare the commentary generated by TennisRAM with the ground-truth commentary and TennisExpert. Red text highlights key descriptions of rally dynamics and decisive events. Across these examples, TennisRAM more closely reflects the reference descriptions of initiative changes and sustained pressure, connecting these processes to specific strokes and point outcomes.

In (a), TennisRAM captures the defense-to-attack transition ("turns defense into attack") and identifies the decisive backhand down the line as a winner, consistent with the reference. In contrast, TennisExpert describes the point as ending with a defensive slice into the net. In (b), TennisRAM emphasizes disruption of the opponent's rhythm before an unforced error, aligning with the reference's description of "change the tempo." In (c), the consistently predicted far-side advantage and depth-pressure rationales are reflected in "keeps the pressure on," linking sustained pressure and side-to-side play to the eventual error. TennisExpert instead focuses on the baseline exchange and the final missed shot, leaving this pressure-building process less explicit. Together, these examples qualitatively support the value of relative advantage modeling for connecting local events to rally progression, yielding commentary that better captures how the point develops rather than merely describing isolated actions or the final outcome.

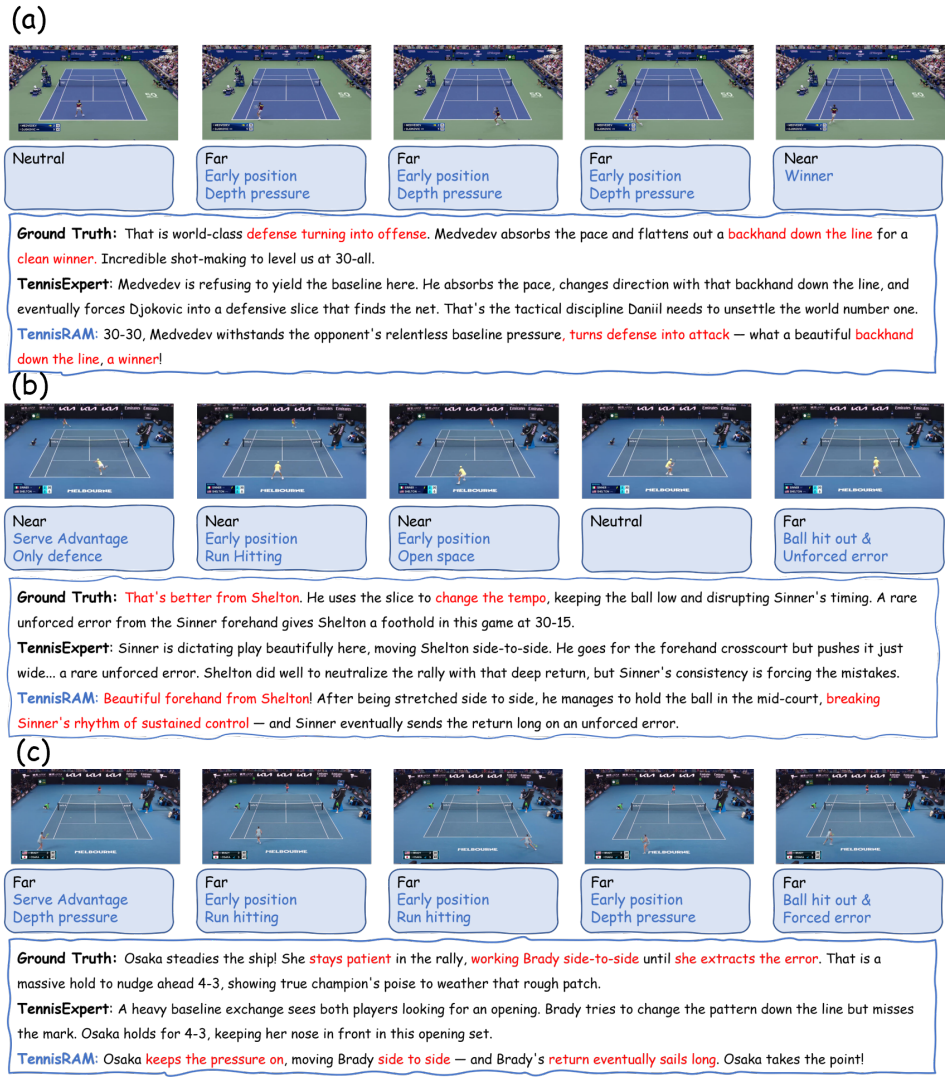

Figure 4: Qualitative comparison on three tennis rallies.