Sportify: Question Answering with Embedded Visualizations and Personified Narratives for Sports Video

Chunggi Lee, Tica Lin, Hanspeter Pfister, and Chen Zhu-Tian

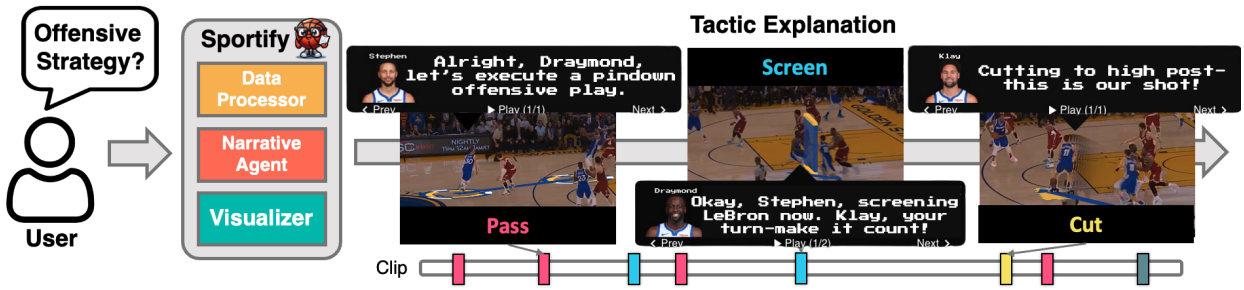

Fig. 1: Sportify explains tactic questions in each clip for everyone, aiming to engage users and foster a love for sports. We integrate embedded visualization and personified narratives generated by large language model (LLM) to elucidate a complex series of actions through action detection, tactic classifier, and LLM pipelines.

Abstract—As basketball's popularity surges, fans often find themselves confused and overwhelmed by the rapid game pace and complexity. Basketball tactics, involving a complex series of actions, require substantial knowledge to be fully understood. This complexity leads to a need for additional information and explanation, which can distract fans from the game. To tackle these challenges, we present Sportify, a Visual Question Answering system that integrates narratives and embedded visualization for demystifying basketball tactical questions, aiding fans in understanding various game aspects. We propose three novel action visualizations (i.e., Pass, Cut, and Screen) to demonstrate critical action sequences. To explain the reasoning and logic behind players' actions, we leverage a large-language model (LLM) to generate narratives. We adopt a storytelling approach for complex scenarios from both first and third-person perspectives, integrating action visualizations. We evaluated Sportify with basketball fans to investigate its impact on understanding of tactics, and how different personal perspectives of narratives impact the understanding of complex tactic with action visualizations. Our evaluation with basketball fans demonstrates Sportify's capability to deepen tactical insights and amplify the viewing experience. Furthermore, third-person narration assists people in getting in-depth game explanations while first-person narration enhances fans' game engagement.

Index Terms—Embedded Visualization, Narrative and storytelling, Basketball tactic, Question-answering (QA) system

1 Introduction

Basketball attracts 400 millions fans worldwide [5], with the NBA Finals alone drawing a peak audience of 17 million [1]. Despite widespread interest in basketball, the rapid pace and intricate dynamics of the basketball plays often leave fans confused and eager for a deeper understanding of the game [31, 66]. Commentary, while helpful, often lacks in providing the abundance of information fans desire, from player performance metrics to complex tactical decisions. Particularly, understanding the *tactics* represents a significant but challenging task. It involves *a series of actions*—screening, passing, cutting, and shooting—that requires significant knowledge to be fully understood [47]. These tactics are crucial in maximizing scoring opportunities, from single plays to team-wide tactics [26, 35, 48]. Players constantly make split-second decisions to either take the shot or distribute the ball, optimizing their team's offensive tactics [49]. A deeper comprehension of these tactics not only enhances the watching experience but also deepens fans' engagement of the game.

Recent efforts in bridging the knowledge gap for fans have embraced the concept of *embedded visualizations* [58]. These visualizations enrich the watching experience by seamlessly integrating data insights within the physical context of the game. Therefore, embedded visualizations have been widely used by both commercial [4, 9, 12] and research systems [31, 66–68] to create *augmented sports videos*. Current augmented sports video products and systems are innovative but limited to deploying predefined visualizations, offering a fixed set of insights for the game's inherently dynamic character [31, 66]. Notably, these tools lack interactive features that would allow fans to investigate the game's tactical dimensions, a limitation that restricts personalized engagement and understanding of complex tactics [31]. This reveals the requirement for advanced game watching systems that enable fans to dynamically explore and query game tactics.

In this paper, we develop a novel Visual Question Answering (VQA) system, Sportify (Figure 1), which integrates embedded visualization and personified narratives to explain basketball tactics. As a VQA system, Sportify focuses on understanding and answering questions about input images, rather than generating visual answers [20]. The design of Sportify is guided by three design considerations thereby enhancing the explanations it provides: 1) video-based tactic explanations, 2) narrative tactic explanations, and 3) embedded visual tactic explanations. Together, these design considerations are crafted to ensure that the generated tactic explanations are reliable, understandable, and engaging for the users.

To answer questions about tactics, Sportify leverages a machine learning pipeline to identify four actions (i.e., Pass, Cut, Screen,

- *Chunggi Lee, Tica Lin, and Hanspeter Pfister are with John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA. E-mail: chunggi_lee, mlin, pfister@g.harvard.edu*
- *Chen Zhu-Tian is with CSE department, University of Minnesota, Minneapolis, MN. E-mail: ztchen@umn.edu. The work was partially done when he was with Harvard.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

and SHOOT), which are critical to understanding basketball tactics, by leveraging the on-court positions of the players. A K-Nearest Neighbors (KNN) algorithm classifies the tactics based on the players' coordinates. The results are subsequently used to generate answers about the logic behind players' actions by using a large language model (LLM). To present the answers, we propose three new action visualizations for three key actions (i.e., PASS, CUT, and SCREEN) and two personified narratives (i.e., **first-person** and **third-person** perspectives), explaining the tactics in a storytelling format. By this, Sportify transforms the passive viewing into an active exploration experiences, offering a deeper understanding and engagement with the game.

We conducted two user studies (i.e., a comparative study and a exploratory study) with basketball fans to investigate the impact on understanding of tactics and in-game decisions. We compare three different conditions and two narrative perspective to explore their impacts on the comprehension of basketball tactics. The results demonstrate that Sportify helps users understand tactics and in-game decisions better, and enhances user engagement and experience compared to existing tactic explanation videos. Furthermore, third-person perspective narration assists people in getting in-depth game explanations while first-person narration enhances fans' game fan and engagement.

In summary, our contributions are as follow: 1) the implementation and design of a novel VQA system for answering tactics and in-game decisions, 2) three novel visualizations such as Cut, Screen, and Pass, integrating the personified narratives (i.e., first or third-person perspective) to provide a storytelling experience that enhances the engagement for fans, and 3) two user studies that evaluate three different conditions with the two personified perspectives texts, and the usability of Sportify comparing to existing tactic explanation videos.

2 RELATED WORK

2.1 Visual Analytics in Sports

Sports data are intrinsically spatial and dynamic. To support analyzing complex sports data for enhancing game understanding and performance evaluation, visualization researchers have developed novel visualization techniques for different sports data and target users.

In particular, basketball has inspired a plethora of novel visualization designs with its intricate interplay among team members and the detailed, organized spatiotemporal data. Targeting sports analysts, BKViz [33] designed an interactive visual analytic system for analyzing individual player performance and team dynamics in a basketball game. Users can analyze heterogeneous data with novel visualizations to support finding patterns and correlations between player actions and performance, such as play-by-play data on a court diagram and player interaction in an arc diagram. OBTracker [59] focused on evaluating the contribution of the player's off-ball movement and presented the player action type and performance in a novel glyph and Voronoi diagram design. HoopInSight [22] compared players' shooting performances using side-by-side shot heat maps and aggregated spatial data presented as location-based glyphs. Other popular sports also attract much attention in the visualization community, including soccer [40, 50], baseball [21], table tennis [56, 57], and badminton [17, 30]. These studies contribute novel visualization approaches to enhance spatiotemporal data analysis and communications in their respective sports.

Targeting non-data experts, some work focused on novel interaction and visualization techniques to support seamless analytic workflow. To support analyzing data during live game viewing, GameViews [64] presents box score views and a game flow chart with key events, along with a chat feature for basketball fans to analyze and discuss game insights live. GameBot [65] proposed using the conversational interface for fans to retrieve game-related data during live basketball games instantly and designed mobile visualizations to enhance data understanding. Lin et al. [31] proposed an embedded visualization framework for analyzing game data within the game context in the basketball game view without the need to switch contexts. More recently, immersive technologies were used to enhance interactive data analysis for coaches and players in racket sports. TIVEE [17] designed an interactive VR interface with embodied interaction to allow analyzing badminton trajectory data in an overview of small multiples and a live-sized badminton court view. VIRD [30] further developed a 3D reconstructed game view based on 2D badminton game videos to support deeper insight analysis for high-performance coaching. Our study builds upon the rich work in sports visualizations and develops novel action-based visualizations for game tactics and in-game decision-making in sports videos.

2.2 Embedded Visualization in Sports Videos

With advanced computer vision techniques, recent research focused on designing visualizations that are directly embedded into sports videos to enhance game analysis of dynamic sports movement.

Stein et al. [50] developed a visual analytic system that combines soccer game videos with trajectory visualizations, applying computer vision methods to derive trajectory measures from the video inputs. Their results show that this embedded visualization method enables expert analysts to perform effective contextualized analysis on team performance. Zhu-Tian et al. [68] proposed a direct manipulation user interface to allow a direct link of the game data to a selected player and contributed a design framework for embedding visual elements with video effects into sports videos, which supports presenting data insights in the sports videos effectively. Lin et al. [31] further tackled the problem of dynamic data requirements throughout sports games by proposing a context-driven embedded visualization framework for live game analysis of sports fans. Yao et al. [60] examined the challenges and design considerations for embedded visualizations in swimming videos from the designer's perspective and found motion context has an impact on the visualization design choices. Zhu-Tian et al. [66] designed gaze interaction to moderate the visualizations shown in basketball videos to avoid visual clutter and enhance fans' game understanding with adaptive visualizations.

Based on the prior research, we identified two gaps in utilizing embedded visualizations within sports videos to enhance fans' game comprehension. First, existing studies predominantly focus on presenting metadata and game statistics (such as athlete names and speeds in [60]) and spatial information (like trajectories and zones in [50]). Aspects that involve more complex data, like game tactics and in-game decision-making, were only preliminary explored by Zhu-Tian et al. [68], which did not target sports fans. Second, there is a lack of customized interactions to support users in retrieving data and analyzing game context during the game. Present approaches tend to offer limited engagement options, such as passive viewing (e.g., gaze in [66]) or basic voice commands [31], which fall short of meeting fans' expectations and allowing more complex data to be explored. Our study addresses this gap by proposing a visual question-answering system that allows fans to explore complex game tactics within sports videos through active conversation and embedded visualizations tailored for action-rich game contexts.

2.3 Visual Storytelling for Spatiotemporal Data

Visual storytelling, especially in the context of spatiotemporal data, has proven to be an effective tactic for communicating complex data [45]. Research grounded in cognitive science has demonstrated that integrating visual and verbal elements enables the construction of a cohesive mental model of a narrative, thereby enhancing the comprehension of complex data [43]. Mayr et al. [34] investigated the organization of temporal and spatial information in supporting narrative comprehension and identified five hybrid visualization techniques, including multiple coordinated views [41], animations [28], layer superimposition [25], layer juxtaposition (or data comics [63]), and space-time cube [27]. As each technique has strengths and limitations in conveying narrative data, it is important to make careful design choices to assist users' internal representation in linking this multimodal information.

Drawing upon this visual storytelling framework to convey complex tactics and decisions in sports, two areas of research are of interest: visual representation of spatiotemporal data and verbal narrative techniques. Several hybrid visualization techniques were adopted for dynamic sports data, including multiple coordinated views [64], layer superimposition [40], and animations [68]. SoccerStories [40] visualizes the sequence of actions in a soccer play with trajectories and linked

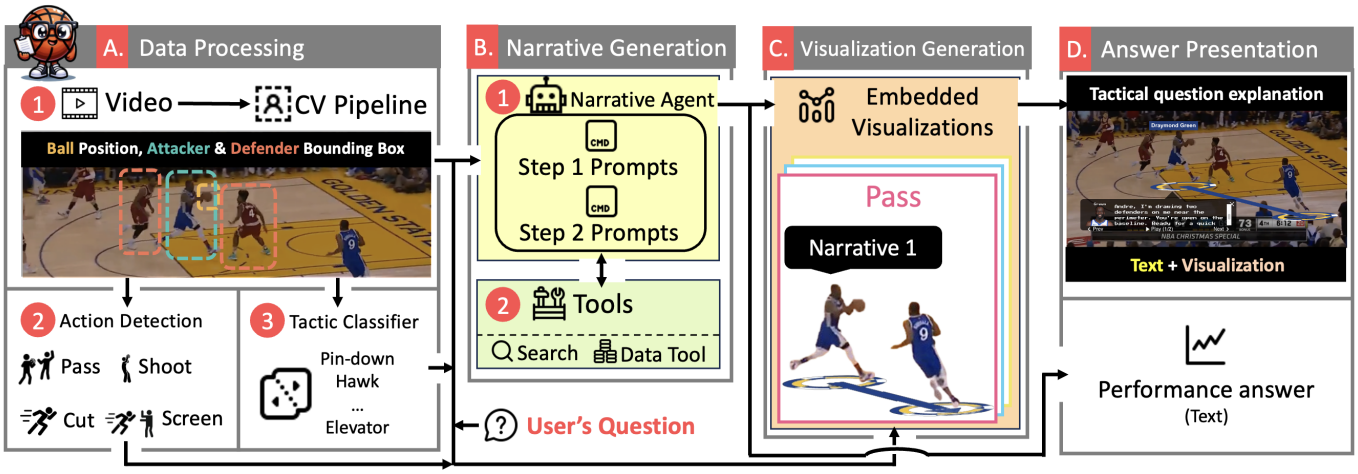

Fig. 2: The pipeline efficiently addresses both tactical and performance-based questions. It begins with data processing (A-1), where videos undergo a computer vision (CV) pipeline to identify players' coordinates, bounding boxes, and the ball's location. This information feeds into action detection (A-2) and tactic classification (A-3), generating tactical textual information for the narrative agent (B-1). Player coordinates and LLM responses are visually embedded (C) and displayed in the video (D). Performance-related queries are handled by the LLM, which retrieves data to provide text-based answers (D).

faceted views on a court diagram. They further proposed a small multiple technique that embeds these diagrams in sports articles to support journalists in communicating data insights in stories. VisCommentator [68] proposed a framework for embedding animated visualizations in sports videos, supporting the creation of data videos with narrative structures like linear and flashback.

In addition to organizing spatial and temporal data views, it is important to explain causality in the sequence of actions for sports game tactics. As shown by Choudhry et al. [16], natural language narratives can complement visualizations when explaining complex causality in network data. Despite different data types, the causal sequences in sports actions share similarities. Zhu-Tian et al. [67] integrated natural language commentary with visual animation in racket sports videos, allowing coupling narratives with animated embedded visualizations to explain game actions in more detail. Building upon the existing work, our work explores using a natural language approach to construct data storytelling for complex sports tactics in basketball. Our novelty lies in adopting LLM-based question-answering techniques to create textual narratives, coupled with animated embedded visualizations in sports videos to create personalized visual storytelling. We also explore how different narrative perspectives, including first- and third-person, impact data comprehension and engagement. To the best of our knowledge, this is the first initiative to blend LLM-generated text narratives with question-answering for sports visual storytelling.

3 Designing Sportify

This section first describes the design considerations behind Sportify and then overviews its three major components.

3.1 Design Considerations

Previous works [31, 66] identified fans are curious and eager to understand the tactics and in-game decisions (e.g., *"the usage of this particular play"* and *"understand the offensive and defensive strategies"*). Thus, our QA system specifically focuses on answering questions related to basketball tactics. Yet, this presents significant challenges, as the answers should *explain a series of complex actions* and presenting the logic behind player movements in an *reliable*, *understandable*, and *engaging* manner. We conceived the design considerations as follow:

R1. Reliable – Explaining Tactics with Grounded Video Data. Ensuring accuracy and reliability in explanations is a critical requirement for QA systems, particularly those analyzing video content. The alignment between the video content and the provided explanations is essential, as the mismatches between the video content and the provided explanations can lead to significant user confusion. This necessitates a mechanism to understand the video and extract data from the video, such as tactic types and the actions involved. This data then serves as the foundation for the QA system to generate explanations. By ensuring that our explanations are directly tied to the observable tactics and actions in the video, we provide users with insights that are not only precise but also verifiable, enhancing the reliability of the information presented by Sportify.

R2. Understandable - Explaining Tactics with Narratives. Storytelling is fundamental in organizing and conveying human experiences, playing a crucial role in how we understand and interpret events [34, 44]. To help users easily understand the tactics employed by teams, we propose the use of well-structured *narratives* that adhere to a logical sequence. This approach not only clarifies the sequence of actions but also reveals the underlying reasons and objectives guiding the players' movements [16]. Moreover, the choice of narrative perspective (i.e., first, second, or third person) also demands careful consideration, as it significantly impacts a viewer's engagement and immersion [14]. In basketball, the third-person perspective aligns with a commentator's view, offering a broad overview of the game, while the first-person perspective resonates with the individual player's decision-making process. Each perspective offers potential benefits in game understanding. Finally, it is essential that these narratives are not only presented as plain text but organized in a structural format that can be effectively mapped or linked with visualizations embedded within the video.

R3. Engaging – Explaining Tactics using Embedded Visualizations. To enrich the explanation of tactics, it is essential to complement the narratives with visual representations, creating an experience akin to watching a film [23]. This requires the careful design of embedded visualizations for the key actions within the narratives. Each action demands specific animated embedded visualizations to capture its unique objectives and to dynamic nature. Furthermore, the visual explanation must also reflect the chosen narrative perspective. For instance, in a first-person narrative, the narration should directly connect to specific subjects within the visualization. In contrast, a third-person narrative allows for a more generalized correlation between the visualizations and the narration. The ultimate challenge for Sportify is to seamlessly integrate these visual explanations within the video content, ensuring a cohesive and engaging presentation of tactical explanation.

3.2 System Overview

Based on the considerations, we have developed Sportify, a visual QA system answering questions about videos [20] and comprising three major components: a Data Processor (Figure 2 a), a Narrative Agent (Figure 2 b), and a Visualizer (Figure 2 c).

At the heart of Sportify lies the Narrative Agent, which leverages

a LLM to interpret the user's question and generate explanations in response. For a system designed for basketball videos, the capability to understand video content is indispensable. Although multi-modal LLMs are capable of processing image data, they often underperform in domain-specific tasks and require a tremendous computation costs, such as detecting actions or tactics from a sports video. To overcome this challenge, our methodology employs a text-only LLM, enriched through the integration of a Retrieval-Augmented Generation (RAG) framework [29] and a Reasoning-and-Actioning (ReAct) prompting strategy [61] for different types of questions. Importantly, Sportify leverages the data extracted from the video as the context (R1) to generate the explanation in a narrative format (R2). These extracted data and explanation are then presented as visualizations embedded in the video (R3). In the subsequent sections, we delve into the specific design and implementation of each component.

4 Data Processor

To create reliable explanations (R1), extracting data from video clips as context for the LLM is crucial. We detail our data extraction method, focusing on identifying tactics and actions, by leveraging the 3D coordinates from the publicly available SportsVU dataset [6] and applying machine learning to determine players and ball positions, following prior work [66]. The 3D coordinates is tracked through multi-camera tracking systems [10], a technique widely adopted in professional basketball leagues, including the NBA [7]. In this study, we utilize the 3D coordinates and extracted data from videos to identify tactics and actions for effectiveness and applicability.

4.1 Tactic Detection

Identifying the tactics is crucial for answering tactic-based questions. We classify a video clip's tactic by comparing five offensive players' movement patterns to those in a reference dataset [53] that includes 134 annotated clips. Each of these clips contains the temporal sequences of the five offensive players' coordinates and the associated offensive tactic, such as Back-Side Pick and Roll, Elevator, or Pin-Down. These patterns are represented by the five temporal sequence of their coordinates, each denoted as $\{(x,y)\}_{i=0}^{t}$, with x and y are the on-court position, and t marking the clip's end frame. Specifically, we leverage the K-Nearest Neighbors (KNN) algorithm and identify the closest match in Figure 2 (A-3). The tactic type of the best match clip in this dataset is assigned to the current video clip in question.

Given the difference in sequence length between video clips, it is impractical to use the traditional Euclidean distance for the KNN algorithm's distance metric. Consequently, we adopt FastDTW, a refined variant of Dynamic Time Warping (DTW), to overcome this challenge. FastDTW [42] can calculate the similarity distance between two temporal sequences of different lengths, thus offering a robust solution for our need. We achieved an accuracy of 85.33%. For more details, see the supplementary material, Section A.

4.2 Action Detection and Filtering

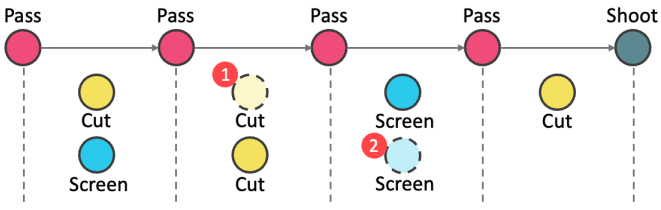

Fig. 3: An action list displays the series of actions performed by offensive players, including Pass, Cut, Screen, and Shoot. The primary actions are identified based on ball movement or movements that enhance scoring opportunities, such as Shoot or Pass the ball. To extract the primary actions related to Pass and Shoot, we set criteria to filter out secondary actions like Cuts and Screens, identified by actions 1 and 2 in red circles.

Action Detection. According to previous work [47], a tactic consists of a series of actions, each associated with a player. Thus, besides the tactic type, we also need to detect the involved actions. In this work, we focus on four most important actions in a basketball offensive tactic in Figure 2 (A-2).:

- Pass occurs when a player transfers the ball to a teammate, who is in a more advantageous position to score. A Pass is detected by tracking *ball ownership* changes within the same team.
- Cut is performed by a player who doesn't have the ball. The player moves swiftly from one court area to another to either create space or distance from a defender, thus enhancing offensive possibilities [18, 52]. For detection, we segment the court into 10 sub-regions (e.g., key, post, wing), following a taxonomy [55]. Then, a Cut is detected if an offensive players moves at a speed of 6 feet per second or faster to a different sub-region [59].
- Screen refers to an offensive player's attempt to block or slow down a defender, thereby creating space and time for a teammate to move into a more advantageous offensive position or to take a shot [52]. From a technical standpoint, a screen occurs when an offensive player blocks a defender who is closely guarding another offensive player with possession of the ball. Thus, to detect a Screen, we analyze the *distances* between players on the court. A Screen is detected if an offensive player changes their marking or covering player.
- Shoot results in a change of possession regardless of whether it scores. We detect a Shoot action if the ball possession changes between the two teams.

The outputs of action detection are a list of actions, each with its timestamp and the associated player.

Action Filtering. Not all captured actions are relevant to the tactic of a team. Including redundant or non-crucial actions could diminish the user experience by cluttering the presentation with unnecessary details. Therefore, we propose to filtering the actions and keep the *important* ones. A practical approach is to prioritize key actions integral to tactic implementation and scoring. According to Tian et al. [52], Pass plays a central role in executing tactics, while Shoot is the end of a tactic. Therefore, as illustrated in Figure 3, we categorize Pass and Shoot as primary actions, while Cut and Screen as secondary actions. Then, we organize the actions chronologically and establish intervals between consecutive Passs (Figure 3).

Next, we filter out the ineffective secondary actions, including 1) Cuts that are not followed by to a ball receive and 2) Screens that are not positioned to benefit the ball handler or the intended receiver, based on the proximity of players and the location of the screen. For example, if a player does not receive the ball after performing a Cut (Figure 3 1), this action is considered as ineffective and removed from the interval. Similarly, the Screen in Figure 3 (2) is excluded if it does not impact the pass, judged by the distance between where the screen is set and the locations of the ball pass before and after the screen.

After the filtering, we obtain a list of important actions that have direct impact to the outcome of the tactic. Our method achieved an F1 score of 73.93% (Details n the supplementary material, Section A).

4.3 Retrieving External Statistics Data

In addition to the data from the current video clip, we also aim to provide the LLM with external meta data, such as the players basic information, team rankings, and historical performance. To achieve this, Sportify is equipped with a suite of tools designed for both in-game data analysis and external information retrieval, including programming tools like Pandas and external APIs like Google Search API, Wikipedia search API, and Statmuse API [11]. The LLM can utilize these tools to extract necessary external statistics data by using a LLM framework named ReAct [61], which enhances LLMs by enabling them to perform tasks and reason in a manner akin to human problem-solving.

5 Narrative Agent

To facilitate the understanding of the tactic (R2), we aim to generate the explanation in a form of narratives (i.e., story). This section introduce the design of the prompt to achieve this goal.

A dynamic prompt with game Context. Rather than forwarding a user's query directly to the LLM, we enhance the input with extra game context from the current video clip. This includes `Player Information`, detected `Tactics`, and `Actions`, all sourced from the Data Processor (Sec. 4). This approach aligns with the Retrieval-Augmented Generation (RAG) framework, which has been proven effective in domain-specific tasks, enabling the LLM to draw upon a vast array of external knowledge without requiring retraining for specific applications. For each query, the system dynamically loads game context relevant to the video clip and construct the prompt, guiding the LLM to tailor its responses to the clip in question. Below is our prompt template, whose details can be found in the supplementary materials.

Prompt template: Please explain { `user question` } based on the following context:

- { `Player Information` }
- { `Detected Tactics` }
- { `Detected Actions` }

Your explanation should follow the below constraints and formats:

- { `Constraints` }
- { `Format` }

Generating Explanations with Narratives. To enhance the comprehension of explanations, we adopt a top-down approach—presenting a tactic overview before delving into detailed actions. Our pilot experiments indicated that LLMs often struggle with generating comprehensive explanations in a single attempt. To address this, we've divided the task into two separate steps, in line with established best practices [46].

First, we prompt the LLM to produce an overview explanation of the tactics. This overview succinctly summarizes the tactics, enabling fans to grasp the essentials of tactical unfolding. Second, we add prompts to the LLM for generating a detailed, sequential breakdown of the tactics, providing users with action-by-action insights into specific actions and decision-making processes. For each step, we incorporate different `constraints` to help the LLM generate different explanations aligned with each step's purpose.

This two-step approach not only refines the LLM's output by incorporating detailed constraints but also safeguards against the generation of incorrect explanatory formats.

Third person perspective narratives:

"Draymond Green cuts from the top to the key aiming to create a scoring opportunity by disturbing the defense. Stephen Curry passes the ball to Draymond Green to create a better scoring opportunity."

First person perspective narratives:

Draymond Green: *"See that gap opening at the top of the key? I'm cutting there now."*
Stephen Curry: *"Got it, This cut will really put pressure on their defense and open up the floor for us."*

Fig. 4: The same tactics' explanation in various narrative perspectives.

Formatting Explanations with Various Perspectives. The standard narrative perspective of LLM is third-person. However, to generate explanations from a first-person perspective—a task LLMs don't typically perform automatically—we introduce specific prompts that steer the model towards generating such responses. Specifically, we prompt the model to frame its explanations as if part of a conversation or role-play dialogue between two players (e.g., ... *answer should be a format of a conversation or a role-play dialogue between two players...*), incorporating descriptions that evoke a first-person reaction to in-game actions (e.g., ... *screen elicit a surprised or shocked reaction from the other person...*). An example format is provided to ensure answer consistency, given the challenge of maintaining uniform response formats due to the varied nature of actions and tactics. Such a prompt technique is applied to both the two aforementioned steps. Figure 4 shows an example of the explanation of the same tactics in different narrative perspectives. Finally, we also prompt the LLM to output the explanations in a structure format to facilitate their mapping to visualizations (Sec. 6).

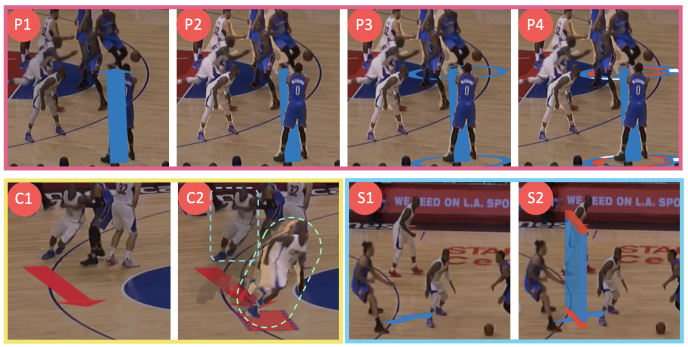

Fig. 5: The iteration design process to design action visualizations (i.e., Pass, Cut, and Screen). From P1 to P4, we remove the occlusion and highlight the two players who send and receive the ball. For the cut, we indicate the exact location that a player will move with flash-forward animation from C1 to C2, while the screen demonstrate a wall to be easily identified a player set on screen from S1 to S2.

6 Visualizing Tactic Explanations

Rather than conveying explanations solely through text, we aim for visually representing them (R3). Since a tactic consists of a series of actions, we decide to visualize these actions to illustrate the explanation of the tactic. Additionally, it's essential to depict the narrative perspectives visually. In the following sections, we will detail the design approach for each of these visual designs.

6.1 Visualizing Actions

Given that the Shoot action signifies the end of a possession, we focus on visualizing the other three actions: Pass, Cut, and Screen. For clearer comprehension, we pause the video during these visualizations. The required data for rendering these visualizations, such as players' coordinates, bounding boxes, and specific frames, is obtained using computer vision models (Figure 2 A1).

Visualizing a Pass. To visualize a Pass, we aim to clearly delineate the dynamics of passing by highlighting the players involved in the pass—the sender and receiver—along with the ball's trajectory. As demonstrated in Figure 5(P4), our visual design employs two rotating circles that mark the sender and receiver, respectively, and an arrow beneath the players to indicate the ball's passing direction. Additionally, we incorporate a flash-forward effect [68] to preview the players' subsequent movements.

The development of our visual design was an iterative process, detailed in Figure 5 P1 to P4. Initially, we used a basic arrow to signify the change in ball possession and its direction (P1), but this approach proved problematic as it occludes the players. To counter this, we repositioned the arrow beneath the players on the court (Figure 5 P2). However, the P2 design encountered visibility issues in crowded scenes. Our subsequent iterations focused on enhancing visibility and understanding: we introduced circles beneath the players to emphasize their roles (P3) and further refined the design (P4) by animating the circles and arrow and adding a flash-forward effect for future movements, and employing team-specific colors for easy identification. This iterative approach, culminating in our final design, facilitates an intuitive and engaging visualization for fans to understand a Pass action.

Visualizing a Cut. The purpose of a Cut in basketball is for a player without the ball to move from one part of the court to another, aiming to create space from defenders and find better positions for shooting [18, 52]. We visualize a Cut by using an arrow to indicate the

Fig. 6: The figure (A) shows the third-person perspective narrative like commentaries, whereas the figure (B) demonstrates the first-person perspective by integrating the action visualizations and narratives around the players to make people more engaged and immersive.

player's movement direction and an area visualization to demonstrate the specific court locations they are moving into (Figure 5 C2). Particularly, we incorporate a flash-forward effect, which previews the player's trajectory before the actual movement occurs. In Figure 5 (C2), a blue dashed line represents the player's position before the cut, and a green line indicates the position after the cut, effectively showcasing the movement path.

Our approach to visualizing a CUT was also iteratively refined. The initial design (Figure 5 C1) depicted a CUT with a simple arrow, highlighting the player's direction but failing to convey the precise future location or the movement trajectory. To address these limitations, the refined design (Figure 5 C2) integrates the arrow with area representation and the flash-forward effect. This enhancement not only clarifies the direction and intent behind a player's CUT but also provides viewers with a predictive insight into the player's positioning, enriching the overall understanding of the game's dynamics.

Visualizing a SCREEN. A SCREEN in basketball is executed to impede or slow a defender, thereby creating space and time for the offensive players. This allows the player with the ball to either move into a more advantageous position, dribble past opponents, or take a shot [52]. To visualize a SCREEN, we aim to enable fans to easily identify when and by whom a screen is set against a defender As shown in Figure 5 S2, our design incorporates both an arrow and a depiction of a screen wall. This combination clearly distinguishes the involved players and the location where the screen occurs.

The initial design (Figure 5 S1) utilized an arrow to indicate the screen action. However, this approach proved insufficient for clear identification, as it often blended with other actions and occluded by the presence of multiple players on the court. We thus improved it and concluded to the current design, which ensuring that fans can easily recognize and understand this crucial aspect of a SCREEN

6.2 Visualizing Narratives

Alongside action visualization, we also need to present the narratives (i.e., the textual explanation of each action) together with their perspectives. We initially intended to provide audio using text-to-speech technology. However, due to the time required to generate the speech, we decided to provide it in text form instead. Figure 6 demonstrates how the scenario of Stephen Curry passing to Andre Iguodala can be narrated from different viewpoints—illustrated through third-person (Figure 6 A) and first-person (Figure 6 B) perspectives.

The third-person perspective, akin to a commentator's overview (e.g., "Andre Iguodala cuts from the post to the wing to space the floor..."), is traditionally used to explain basketball plays (Figure 6 A). This narration style fits seamlessly into visualizations using a single chatbox, streamlining integration without additional interface requirements.

Conversely, the first-person perspective, which captures players' emotions and dialogues (e.g., *"Perfect timing, Andre. Cutting towards the high post to create a mismatch..."*), can confuse fans when presented in a traditional chatbox format. To address this, we adopt dialogue bubbles for first-person narratives, akin to comic book styles (Figure 6 B). These bubbles move with the players on screen, enhancing engagement by allowing users to interact with the narrative—navigating through the conversation with 'previous', 'play', and 'next' controls.

7 USER STUDY

We conduct a two-phased user study with basketball fans to evaluate the understanding, usability, and engagement of Sportify. Two experiments were conducted: a comparative study of narratives with first and third-person perspectives with and without visualizations, and an exploratory user study on the overall experience of Sportify.

7.1 Participants & Experiment Set-Up

We recruited 13 basketball fans (P1-P13; M = 13; Age: 23 - 33) via university mailing lists. Due to a technical issue, we were only able to collect P4's subjective feedback, excluding the task completion time and accuracy. Participants reported their fandom levels, including 3 novice, 2 casual fans, and 8 engaged fans. In addition, participants reported their frequency of watching basketball in four different levels: 4 participants watched at least 1 game per week, 5 watched 2-4 games per month, 1 watched 11-23 games per year, and 3 watched 1-10 games per year. We selected two famous games: one between the Golden State Warriors and the Cleveland Cavaliers on December 25th, 2015, and the other between the Oklahoma City Thunder and the Los Angeles Clippers on December 21st, 2015. These games were featured as the best by the NBA in the 2015-16 season. The two experiments were conducted in person on the same day, using a 14-inch laptop for setup. The study took about one hour to finish and all participants were compensated with a $20 gift card.

7.2 Study Design & Measure

Introduction & pre-survey (10 mins). Before starting the user study, we introduced our user study and received the consent form. We collected information about participants' backgrounds, such as fandom level and average watching frequency, through a pre-study survey.

Task 1: A comparative study (25 mins). Our first experiment investigates the understanding of strategies, the helpfulness of visualization and personified narratives. Participants first watched a game clip with an explanation, and were then asked to order a list of actions (e.g., pass, cut, and screen) to match the action sequence in the game based on their understanding. The explanation was delivered in one of the three conditions, including explaining strategies using pure text (*Text*), text with visualization in third-person narrative (*Third*), and text with visualization in first-person narrative (*First*) conditions. The explanations were identical between *Text* and *Third* conditions. Explaining tactics often requires an in-depth understanding and detailed explanation, making the textual narrative necessary [13, 16, 19, 54]. Therefore, the non-text version was excluded from the baseline. We prepared twelve clips from both games evenly and tasked participants with their understanding of the strategy in the game. Each condition has four trials, including a practice and three actual trials, in which we collected task completion time and accuracy. The number and types of actions are evenly distributed among three conditions. In addition, the order of the conditions and the twelve clips were counterbalanced in the study.

During the study, participants were introduced to three conditions and practiced the trials in the training session. In the study, participants watched the assigned game clip and used a drag-and-drop interface to move the order of the list of actions to match the sequence in the

game. After completing all 12 trials, participants ranked their preferred explanation conditions and rated the helpfulness of the visualizations and narratives, providing subjective feedback in the questionnaire.

Task 2: An exploratory study (25 mins). The second part of the user study involves a free exploration of Sportify to compare participants' game-watching experience with tactic explanation videos on YouTube and using Sportify. The participants chose either a first- or third-person perspective based on their preference, and analyzed a game selected from two games used in Task 1 (GSW vs. CLE or OKC vs. LAC) using Sportify. As a baseline for a typical tactic explanation approach for basketball games, we selected three tactic explanation videos from the popular basketball channels on YouTube [2,3,8], watching 2-3 minutes.

In the study, the participants first watched a YouTube video randomly assigned from [2,3,8]. After that, they were introduced to all features of Sportify in a training video. Participants freely watched the game video and asked any questions at any time using Sportify. To inform users when they can ask tactical questions, our system visually indicates which parts of video clips can be questioned and which cannot. Following previous work [31,38,66], we measured user experiences using subjective rating questions in the post-study survey, including "It was helpful", "It was fun", "I felt in control", "I felt encouraged", "I am likely", and "I felt engaged" when using Sportify, and collected feedback from the survey and think-aloud methods during the study.

8 Study Results

We present the results of two tasks in our user study and discuss how visualizations and different narrative perspectives affect the understanding of game strategy. Besides, we discuss how Sportify differs from existing strategy explanation videos in enhancing game understanding.

8.1 How Do Three Different Conditions Affect Understanding Strategy Explanations?

In Task 1, participants matched the sequence of actions in twelve clips under three conditions: 1) pure text in a third-person perspective (*Text*), 2) text with visualization in a third-person perspective (*Third*), and 3) text with visualization in a first-person perspective (*First*). We presented the findings on how visualization and personified narrations affect strategy comprehension, task performance, and overall experiences.

8.1.1 Watching Time Increased Without Affecting Accuracy

We measured the accuracy of nine game clips in the trials from 12 participants, excluding three practice clips. The accuracy of matching the actions to game strategy for the *Text* and *Third* conditions were identical at 72.22% (26 out of 36), while the *First* condition was slightly lower at 69.44% (25 out of 36). The average watching time for the *First* condition (86.27s) exceeded that of the third-person perspective (*Text*: 69.44s and *Third*: 68.72s). However, the average solving time, which measures the time participants took to complete ordering actions in the trial after watching the clip, was similar across all three conditions (*Text*: 40.33s, *Third*: 39.94s, and *First*: 44.53s). After conducting a normality test, we found that none of the conditions followed a normal distribution, so we performed a Kruskal-Wallis test. This test revealed a significant difference in watching time among the three conditions (p= 0.02 and H= 7.765) but not in solving time (p= 0.89 and H=0.24). Dunn's post-hoc test further identified a significant difference between the *Text* and *First* conditions (p=0.046) in watching time.

Overall, the results show that the *Third* condition has a similar watching time to the *Text* condition, with visualizations aiding users in understanding the main actions in the clip. Participants found visualizations in *Third* helpful, as noted *"The explanations with visualizations help more than the text in general (P9)"*. In addition, participants spent 25% more time watching the video with first-person narrative in *First* than the third-person narrative in *Text* and *Third* condition. This is likely due to that the *First* condition involves more interactions than the other two conditions due to the personified narratives presented between a pair of players, leading to longer time to interact and engage in the video. Participants did not find this interaction impeding their understanding, but instead felt that *First person perspective text was more enjoyable and simpler to understand than the third person perspective text (P13)*. *First* condition was found to be most helpful and enhanced their confidence, as shown in Sec. 8.1.2.

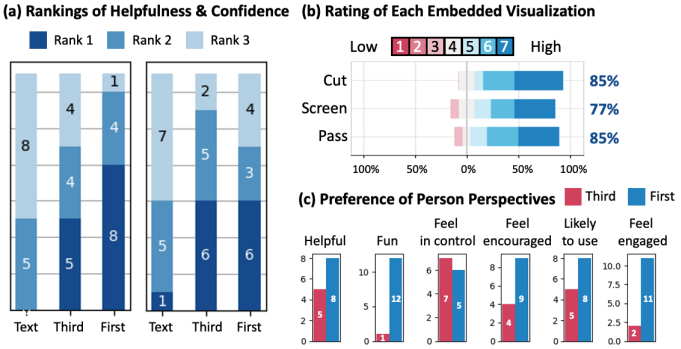

Fig. 7: A task 1 user study results. Figure (a) reveals that the First person perspective ranks highest in helpfulness and confidence across three conditions. Figure (b) indicates positive participant ratings for each embedded visualization's helpfulness. Figure (c) compares usability across two different narrative perspectives.

8.1.2 First-Person Narrative Enhanced Feeling of Helpfulness and Confidence

Figure 7 (a) reveals that the *First* condition was rated highest in helpfulness and enhancing their confidence in understanding by 8 and 6 participants out of 13 participants, respectively, followed by the *Third* condition with 5 and 6 participants, and the *Text* condition with none and 1 participant. Participants highlighted the advantages and drawbacks of the *First* and *Third* conditions. Six participants noted that placing the chat box near players was helpful and not distracting, stating that *"First person perspective chat box comes closer to players, which helps me better understand tactics. (P1)"*. They also appreciated the interactive "Play" button for navigating conversations, feeling an experience similar to playing a video game, with one saying, *"I am 'in control' of the players, almost like a video game."* However, some found the *First* condition confusing due to its abundant text and lack of resemblance to real commentary, with feedback like, *"First condition felt a bit informal and not realistic, (P12)"*.

Four participants preferred the *Third* condition for its realism and detailed information, commenting, *"The Third condition was more effective in conveying detailed information."* However, they reported losing focus due to the need to switch attention between visualizations and text, with one mentioning, *"I was distracted by moving my eyes between players and chat box (P1)"*.

Overall, participants found the *First* and *Third* narrative perspectives more helpful than *Text* alone in enhancing their understanding of game strategies. With these visual narratives, participants felt more confident in grasping game strategies, with participants equally favoring the *First* and *Third* perspectives. The *First* offered interactivity and embedded placement of the visualizations, while the Third provided realism and clearer comprehension.

8.1.3 Embedded Visualizations Help Identifying Player Interactions and Ball Movements

As shown in the Figure 7 (b), all the three visualizations received positive ratings from the majority, including Cut (85%), Screen (77%), and Pass (85%), confirming the usefulness of the three visualizations for understanding strategic explanation.

Participants, especially novice and casual fans, commented that these action-specific visualizations significantly aided their understanding of basketball plays. In particular, visualizations significantly aided in identifying "Screen" and "Cut" actions, allowing participants to comprehend complex plays like off-ball movements and interactions between offensive and defensive players more easily. This result aligns with our main goal of elucidating off-ball movements by pinpointing the screen's location and direction with the Screen visualization and adding

a flash-forward effect to demonstrate the next player move in the Cut visualization. Furthermore, participants found the Pass visualization very helpful for its clear depiction of ball movement directions, senders, and receivers. For instance, P7 mentioned, *"I sometimes miss who receives the pass, but with visualization, I could easily notice it."* This feedback confirms the primary design goal of the Pass visualization in highlighting the players initiating and receiving passes.

However, watching the simultaneous presentation of the Screen visualization and text presents a challenge, as noted by P11: *"It would take some time to get used to seeing the text and visualization, since it shows a lot at once."* This indicates a learning curve for users when simultaneously comprehending the visualization and text integration.

8.1.4 First-Person Enhances Fun and Enjoyment, While Third-Person Allows Feeling of Control and Formality

Figure 7 (c) shows the usability ratings for the *First* and *Third* person perspective narrative designs. This result revealed a clear preference for *First* person perspective narrations over *Third* person on fun and engagement by 12 and 11 participants out of 13. This preference is attributed to the enjoyment of story-based narratives and a user interface design that simulates players engaging in conversation with one another. P2 praised the narrative for its use of nicknames, expressing *"I like the story-based narrative where they even called each other nicknames like Matty"*. The *First* person perspective was also seen as creative and interactive, with P11 finding the player communication innovative and enjoying the conversational navigation. P11 noted that *"I felt most interactive and creative when manipulating players' conversations."* Moreover, participants found this perspective helpful in understanding player thoughts and intentions, enhancing their grasp of overall tactics, as P5 mentioned *"It is easier to understand each player's intention..."* However, P8, a former university team player, criticized the dialogue in this condition for lacking realism and not accurately reflecting players' in-game thought processes. As a result, our results reveal that the first-person perspective enhances engagement, encouragement, and fun, supporting the findings in previous research [14, 37].

On the other hand, Figure 7 (c) demonstrates that from the post-survey feedback, 7 and 5 participants found the *Third* person perspective narratives make them feel in control and found it more helpful, respectively. The text in third-person perspective is viewed as more concise and accurate for strategy comprehension compared to the first-person perspective. P3 stated *"Third person perspective avoids unnecessary information like 'Now, it is time to shoot'."*. Furthermore, some participants favored the third-person perspective for its conventional and formal explanatory style like game commentaries, mirroring their typical basketball viewing experiences. As P6 mentioned, *"Third person perspective seems more standard, formal, and familiar to me."* Additionally, P9 noted that the third-person perspective is easier to follow and control with its predictable layout, highlighting that *"It's easy to lose attention while following the chatbox above the players, and the third-person perspective is easier to control with the static chatbox navigation button."* According to other research [36], shifting the perspective from third-person to first-person requires additional reading time and effort, causing discomfort for users who are accustomed to the third-person view.

In sum, participants' feedback distinctly indicates that first-person narratives provide a greater sense of fun and engagement, inspiring users to explore further, whereas third-person narratives, with their familiar format, enhance the feeling of control.

8.2 How Does Sportify Compare to Traditional Basketball Strategy Explanation Videos?

Task 2 evaluated Sportify's usability by asking participants to watch an existing strategy explanation video and then freely used Sportify.

Participants predominantly favored Sportify over the existing strategy explanation videos, as shown in numbers of participants on metrics of helpful (9), fun (12), feel in control (12), feel encouraged (11), likely to use (11), and feel engaged (10) in Figure 8 (a). The majority of participants valued Sportify's capability to generate personalized answers to their queries and provide infinite content. Notably, two

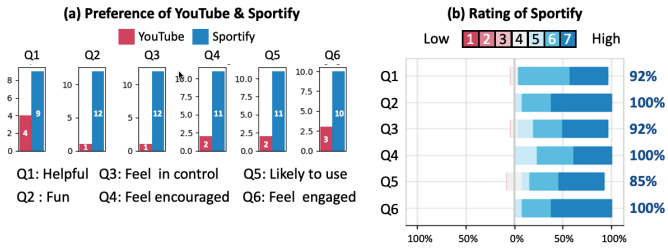

Fig. 8: Sportify received better results compared to existing tactical explanation videos, with positive outcomes in all questionnaire of usability.

casual fans mentioned that while the existing videos left their questions unanswered, Sportify provided the desired information on demand. P5 highlighted *"Sportify's advantage of offering limitless, real-time content compared to the fixed material in traditional videos."* The tool's explicit presentation of strategies and motivations was especially helpful, with one user noting, *"It helped me see the clear motivation and strategy explanation from a first-person view (P4)"*, underscoring Sportify's effectiveness in delivering detailed play insights.

On the other hand, some participants preferred the existing explanation videos for their structured content and insights from professional analysts. Engaged fans with an understanding of basic strategies appreciated the depth of professional analysis, as P4 noted, *"More rigorous analysis of plays with explanations from professional analysts."* Additionally, the advocates of the video highlighted their ease of use, pointing out that they demand less effort to navigate, with P9 stating, *"The video is easier to watch since it requires less effort."*

Figure 8 (b) illustrates Sportify's usability ratings on a scale from 1 (strongly disagree) to 7 (strongly agree), with all aspects receiving positive feedback (above 85%). Sportify received notable 100% positive responses in fun, encouragement, and engagement. Users particularly enjoyed the first-person perspective for its dynamic explanation of tactics and player thought processes, as P4 highlighted: *"It described the thought processes of players in a fun way, which helped me understand."* The ability to ask personalized questions at their own pace encouraged participants to explore Sportify, exploring specific areas of interest and seeking detailed information. P11 describes *"It moves at your own pace, allowing you to probe the system with more specific questions about any play."* The enhanced accessibility and in-depth analysis motivated users to engage more deeply with the strategies, with P12 appreciating the *"in-depth analysis that focuses on overall strategy, especially showing the thought processes and reasoning among players."* P5 commented a strong positive feedback in terms of "likely to use", citing its real-time capabilities as particularly appealing: *"If this system is public, I will absolutely try it. The real-time ability is really cool."* In summary, eleven participants found the system extremely helpful, praising its accurate and sensible explanations and the freedom to inquire at will.

9 Discussions

We summarize the design insights derived from developing our system, along with feedback and observations from the user studies.

9.1 Sports Narratives: Various Preferences for Perspective

Previous studies [24, 37] have shown that readers familiar with the events and emotions of a story's protagonist are more likely to choose a first-person perspective, while those with less experiences prefer a third-person perspective. Based on this, we thought that novice or casual fans less familiar with basketball would choose the third-person perspective, while engaged fans would select the first-person perspective.

However, during free exploration of our system, participants showed an equal preference for both perspectives: seven for first-person and six for third-person. Among the participants, two novices and one casual fan chose the first-person perspective, as did four out of eight engaged fans. This trend was consistent across various levels of basketball knowledge and viewing frequency. Unlike general storytelling, a

crucial factor here is the need to understand complex basketball strategies. As mentioned in [66], novice and casual fans tend to enjoy games more for their entertainment value and likely prefer the first-person perspective for its immersive experience. In contrast, engaged fans focus on technical aspects and strategies, often seeking a deeper understanding, and might lean towards the third-person perspective for its broader overview. Therefore, this mixed conclusion likely stems from individual familiarity with the protagonist and the need for a deeper understanding of the narratives.

9.2 Using Text-Based LLMs Instead of Multi-Modal LLMs

In developing Sportify, we experimented with utilizing current state-of-the-art multi-modal LLMs [15, 32, 62] that are capable of processing images or videos to interpret visual content. However, we found that applying multi-modal LLMs is still challenging in the sports domain, which is characterized by numerous players and complex dynamics and interactions. Instead, we transformed visual data from the videos into textual information using a computer vision pipeline, providing the LLM with detailed information such as actions and tactical insights to facilitate answer generation. In this framework, the LLM served as a central hub or "brain" to synthesize all information and make inferences. In the future, we anticipate that advancements in multi-modal LLMs' capabilities in understanding images and videos will enable a more integrated approach to system development, offering a seamless, all-encompassing solution. One limitation of our study was the inability to provide defensive tactics due to the lack of data, but we believe that this will be possible with multi-modal LLMs in the future.

9.3 Presenting LLM-Generated Text in Videos

Explaining the dynamics of sports is challenging. Previous research has mainly focused on enhancing understanding through embedded visualizations, often neglecting textual descriptions. With text descriptions generated by LLMs, we have explored different methods to present these descriptions in the video.

Presenting Text with Visualizations. Presenting LLM-generated text with visualizations in the video is an intuitive method. However, simply integrating LLM-generated text and visualizations without careful design consideration could hinder the user experience. As identified in our paper, placing third-person perspective text with visualizations forces viewers to switch between action visualizations around players and text in the chat box, leading to a loss of concentration and degrading the user experience. This highlights the need for careful visualization designs and user interfaces. While LLMs can potentially identify positions to place text with visualizations automatically, they are not specifically designed for particular tasks, making it challenging to achieve complex output. Therefore, simplifying the problem through an intermediary step and then subsequent stages such as visualization helps us achieve better results [46, 51].

Presenting Text using Audio. Another alternative we considered was the use of audio. We attempted to convert the generated text into audio and play it together with the visualizations, but found that this process was time-consuming. Consequently, we decided to omit this feature as it could degrade the user experience. However, during our user surveys, two participants inquired about the availability of audio. We believe that incorporating audio with visualizations could potentially allow for a more seamless integration.

9.4 Designing User-Centered LLMs for Domain-Specific Tasks

Strategies for Ensuring Consistent LLM Responses. Given the unfamiliarity of first-person narratives to LLMs, ensuring a uniform style of response presented a significant challenge. The complexity increased when attempting to generate conversation-like formats where multiple players interact and address each other by name. To achieve consistency in responses, we applied multiple strategies:

First, we utilized a template-based approach within the prompts, integrating various constraints as suggested by previous studies [46]. Despite these efforts, the LLM still struggled to deduce the logic or reasoning behind in-game decisions from textual input alone. To aid in this reasoning process, we explicitly outlined the rationale for 3 to 5 actions within the prompt.

Additionally, we employed ReAct (Reason + Act) [61], which generates human-like task-solving trajectories and prevents error propagation. We believe that these approaches help prevent issues of hallucination. However, if hallucinations still occurred despite these methods, users were encouraged to rephrase their questions.

Selecting Key Information for Enhanced LLM Performance. The number of actions detected in a video is usually too many for the LLM to reason through effectively. By filtering these actions to include only essential information for the LLM, we streamlined the data, akin to applying an importance score for selecting critical information [39]. These strategies resulted in a more precise system, surprising users with its accuracy in providing explanations during the user study. Feedback from the study highlights the critical role of fusing domain-specific knowledge to curate information for the LLM in developing LLM applications. Our approach of guiding the LLM with key information proved to be effective and satisfying.

Addressing Various User Needs. We observed that not every fan seeks intricate tactical insights or understands the players' thought processes. Several participants favored succinct explanations, indicating a preference for customizable narrative detail levels. This suggests that allowing users to select their desired level of detail could significantly enhance their experience.

9.5 Limitations

The sample size of our study is comparable to other similar sports visualization papers [31, 66]. Besides, the focus of our study is on qualitative feedback rather than quantitative results. Through qualitative feedback, we found that not only novices but also engaged fans benefited from and were immersed in understanding the tactics. Nonetheless, further experiments with a larger user base are beneficial. In real-world settings, our VQA system faces challenges related to data precision, processing speed, and user diversity. The precision of 3D tracking data from single monocular videos is often insufficient, but this can be improved with 3D vision models using multiple-angle videos or extra sensors. The system's multiple pipelines require significant processing time, hindering real-time performance, though this can be mitigated with better machinery and lighter models. Additionally, customizing conversation content to suit diverse user needs and skill levels, including adjustable explanation detail and length, is crucial. Lastly, our initial focus was on explaining tactics through key actions in the proposed visualizations. However, more diverse visual explanations are needed to fully support the tool's practical use.

10 Conclusion

Our work introduces Sportify, an innovative VQA system. It significantly enriches the basketball watching experience by integrating embedded visualization with personalized narrative explanations of tactics. This novel system enables fans to investigate understanding the game through both statistical queries and complex tactical questions. By employing a computer vision pipeline and leveraging a LLM to generate insightful explanations of players' logic and reasoning, Sportify transforms the paradigm from passive watching to an interactive, engaging exploration of basketball. Our user studies evaluate three different conditions and two narrative perspectives in a comparative study, as well as the usability of our tool in a free exploratory study. The results reveal that Sportify significantly enhances users' comprehension of tactics made during games and elevates user engagement and experience beyond what is offered by existing tactic explanation videos. Moreover, narration from a third-person perspective aids in providing detailed explanations of the game, while first-person perspective increases fans' enjoyment and engagement with the game.

ACKNOWLEDGMENTS

This work is supported by SEAS Graduate Fellowship, NSF grant IIS-1901030, NIH grant R01HD104969, NSF grant III-2107328.

REFERENCES

[1] 23 amazing nba viewership statistics in 2024. "https://playtoday.co/blog/stats/nba-viewership-statistics/". Accessed on March 25, 2024. 1

[2] 5 clever nba set plays and strategies explained. "https://www.youtube.com/watch?v=Fd3MzuHKHHI". Accessed on March 21, 2024. 7

[3] 6 genius nba plays explained. "https://www.youtube.com/watch?v=lpR9Fp84XPw&t=146s". Accessed on March 21, 2024. 7

[4] Court vision. "https://www.clipperscourtvision.com/". Accessed on March 25, 2024. 1

[5] The most popular sports in the world. "https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html". Accessed on March 25, 2024. 1

[6] Nba sportvu dataset. "https://paperswithcode.com/dataset/nba-sportvu". Accessed on March 21, 2024. 4

[7] Nba website. "https://www.nba.com/". Accessed on March 25, 2024. 4

[8] One of my favorite nba offensive concepts. "https://www.youtube.com/watch?v=_wA4Fpzx08s". Accessed on March 21, 2024. 7

[9] Second spectrum. "https://www.secondspectrum.com/". Accessed on March 25, 2024. 1

[10] Sportvu camera system in nba. "https://www.statsperform.com/team-performance/basketball/optical-tracking/". Accessed on March 25, 2024. 4

[11] Statmuse. "https://www.statmuse.com/". Accessed on March 14, 2024. 4

[12] Viz libero. "https://www.vizrt.com/products/viz-libero.". Accessed on March 25, 2024. 1

[13] G. Altavilla, G. Raiola, et al. Global vision to understand the game situations in modern basketball. *Journal of Physical Education and Sport*, 14:493–496, 2014. 6

[14] M. Chen and R. Bunescu. Changing the narrative perspective: From deictic to anaphoric point of view. *Information Processing & Management*, 58(4):102559, 2021. 3, 8

[15] W.-G. Chen, I. Spiridonova, J. Yang, J. Gao, and C. Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*, 2023. 9

[16] A. Choudhry, M. Sharma, P. Chundury, T. Kapler, D. W. Gray, N. Ramakrishnan, and N. Elmqvist. Once upon a time in visualization: Understanding the use of textual narratives for causality. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1332–1342, 2020. 3, 6

[17] X. Chu, X. Xie, S. Ye, H. Lu, H. Xiao, Z. Yuan, C. Zhu-Tian, H. Zhang, and Y. Wu. TIVEE: Visual Exploration and Explanation of Badminton Tactics in Immersive Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 2021. 2

[18] J. Courel-Ibáñez, A. P. McRobert, E. Ortega Toro, and D. Cárdenas Vélez. Inside game effectiveness in nba basketball: Analysis of collective interactions. *Kinesiology*, 50(2.):218–227, 2018. 4, 5

[19] J. Courel-Ibáñez, A. P. McRobert, E. O. Toro, and D. C. Vélez. Collective behaviour in basketball: a systematic review. *International Journal of Performance Analysis in Sport*, 17(1-2):44–64, 2017. 6

[20] A. C. A. M. de Faria, F. d. C. Bastos, J. V. N. A. da Silva, V. L. Fabris, V. d. S. Uchoa, D. G. d. A. Neto, and C. F. G. d. Santos. Visual question answering: A survey on techniques and common trends in recent literature. *arXiv preprint arXiv:2305.11033*, 2023. 1, 3

[21] C. A. Dietrich, D. Koop, H. T. Vo, and C. T. Silva. Baseball4d: A tool for baseball game reconstruction & visualization. *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 23–32, 2014. 2

[22] Y. Fu and J. T. Stasko. Hoopinsight: Analyzing and comparing basketball shooting performance through visualization. *IEEE Transactions on Visualization and Computer Graphics*, 30:858–868, 2023. 2

[23] N. Gershon and W. Page. What storytelling can do for information visualization. *Communications of the ACM*, 44(8):31–37, 2001. 3

[24] M. C. Green. Transportation into narrative worlds: The role of prior knowledge and perceived realism. *Discourse processes*, 38(2):247–266, 2004. 8

[25] W. Javed and N. Elmqvist. Exploring the design space of composite visualization. In *2012 ieee pacific visualization symposium*, pp. 1–8. IEEE, 2012. 2

[26] I. S. Kohli. On optimal offensive strategies in basketball. *arXiv preprint arXiv:1506.06687*, 2015. 1

[27] M.-J. Kraak. The space-time cube revisited from a geovisualization perspective. In *Proc. 21st international cartographic conference*, pp. 1988–1996. Citeseer, 2003. 2

[28] S. Kriglstein, M. Pohl, and M. Smuc. Pep up your time machine: Recommendations for the design of information visualizations of time-dependent data. *Handbook of human centric visualization*, pp. 203–225, 2014. 2

[29] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 4

[30] T. Lin, A. Aouididi, C. Zhu-Tian, J. Beyer, H. Pfister, and J.-H. Wang. VIRD: Immersive Match Video Analysis for High-Performance Badminton Coaching. *IEEE Transactions on Visualization and Computer Graphics*, 30:458–468, 2023. 2

[31] T. Lin, C. Zhu-Tian, Y. Yang, D. Chiappalupi, J. Beyer, and H. Pfister. The quest for omnioculars: Embedded visualization for augmenting basketball game viewing experiences. *IEEE transactions on visualization and computer graphics*, 29(1):962–971, 2022. 1, 2, 3, 7, 9

[32] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 9

[33] A. G. Losada, R. Therón, and A. Benito. Bkviz: A basketball visual analysis tool. *IEEE Computer Graphics and Applications*, 36:58–68, 2016. 2

[34] E. Mayr and F. Windhager. Once upon a spacetime: Visual storytelling in cognitive and geotemporal information spaces. *ISPRS International Journal of Geo-Information*, 7(3):96, 2018. 2, 3

[35] A. McIntyre, J. Brooks, J. Guttag, and J. Wiens. Recognizing and analyzing ball screen defense in the nba. In *Proceedings of the MIT sloan sports analytics conference, Boston, MA, USA*, pp. 11–12, 2016. 1

[36] D. S. Miall and D. Kuiken. Shifting perspectives: Readers' feelings and literary response. *New perspectives on narrative perspective*, pp. 289–301, 2001. 8

[37] M. Mulcahy and B. Gouldthorp. Positioning the reader: the effect of narrative point-of-view and familiarity of experience on situation model construction. *Language and Cognition*, 8(1):96–123, 2016. 8

[38] H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, 2010. 7

[39] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. 9

[40] C. Perin, R. Vuillemot, and J.-D. Fekete. Soccerstories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, 19:2506–2515, 2013. 2

[41] J. C. Roberts. Exploratory visualization with multiple linked views. In *Exploring geovisualization*, pp. 159–180. Elsevier, 2005. 2

[42] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007. 4

[43] W. Schnotz. An integrated model of text and picture comprehension. *The Cambridge handbook of multimedia learning*, 49(2005):69, 2005. 2

[44] K. Schröder, W. Eberhardt, P. Belavadi, B. Ajdadilish, N. van Haften, E. Overes, T. Brouns, and A. C. Valdez. Telling stories with data–a systematic review. *arXiv preprint arXiv:2312.01164*, 2023. 3

[45] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010. 2

[46] L. Shen, Y. Zhang, H. Zhang, and Y. Wang. Data player: Automatic generation of data videos with narration-animation interplay. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 5, 9

[47] A. Sicilia, K. Pelechrinis, and K. Goldsberry. Deephoops: Evaluating micro-actions in basketball using deep feature representations of spatio-temporal data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2096–2104, 2019. 1, 4

[48] A. Singh. Optimizing performance in basketball: A game-theoretic approach to shot percentage distribution in a team. *arXiv e-prints*, pp. arXiv–2310, 2023. 1

[49] B. Skinner and M. Goldman. Optimal strategy in basketball. In *Handbook of statistical methods and analyses in sports*, pp. 245–260. Chapman and Hall/CRC, 2017. 1

[50] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlücke, T. Schreck, G. L. Andrienko, M. Grossniklaus, and D. A.

Keim. Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24:13–22, 2018. 2

[51] N. Sultanum and A. Srinivasan. Datatales: Investigating the use of large language models for authoring data-driven articles. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 231–235. IEEE, 2023. 9

[52] C. Tian, V. De Silva, M. Caine, and S. Swanson. Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. *Applied Sciences*, 10(1):24, 2019. 4, 5, 6

[53] T.-Y. Tsai, Y.-Y. Lin, H.-Y. M. Liao, and S.-K. Jeng. Recognizing offensive tactics in broadcast basketball videos via key player detection. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 880–884. IEEE, 2017. 4, 12

[54] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International journal of human-computer studies*, 57(4):247–262, 2002. 6

[55] Unkown. Glossary of basketball terms. "https://en.wikipedia.org/wiki/Glossary_of_basketball_terms", Oct. 2010. Accessed on March 14, 2024. 4

[56] J. Wang, J. Wu, A. Cao, Z. Zhou, H. Zhang, and Y. Wu. Tac-miner: Visual tactic mining for multiple table tennis matches. *IEEE Transactions on Visualization and Computer Graphics*, 27:2770–2782, 2021. 2

[57] J. Wang, K. Zhao, D. Deng, A. Cao, X. Xie, Z. Zhou, H. Zhang, and Y. Wu. Tac-simur: Tactic-based simulative visual analytics of table tennis. *IEEE Transactions on Visualization and Computer Graphics*, 26:407–417, 2020. 2

[58] W. Willett, Y. Jansen, and P. Dragicevic. Embedded data representations. *IEEE transactions on visualization and computer graphics*, 23(1):461–470, 2016. 1

[59] Y. Wu, D. Deng, X. Xie, M. He, J. Xu, H. Zhang, H. Zhang, and Y. Wu. Obtracker: Visual analytics of off-ball movements in basketball. *IEEE Transactions on Visualization and Computer Graphics*, 29:929–939, 2022. 2, 4

[60] L. Yao, R. Vuillemot, A. Bezerianos, and P. Isenberg. Designing for visualization in motion: Embedding visualizations in swimming videos. *IEEE Transactions on Visualization and Computer Graphics*, 30:1821–1836, 2023. 2

[61] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. 4, 9

[62] S. Zhang, P. Sun, S. Chen, M. Xiao, W. Shao, W. Zhang, K. Chen, and P. Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 9

[63] Z. Zhao, R. Marr, and N. Elmqvist. Data comics: Sequential art for data-driven storytelling. *tech. report*, 2015. 2

[64] Q. Zhi, S. Lin, P. T. Sukumar, and R. A. Metoyer. Gameviews: Understanding and supporting data-driven sports storytelling. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. 2

[65] Q. Zhi and R. A. Metoyer. Gamebot: A visualization-augmented chatbot for sports game. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. 2

[66] C. Zhu-Tian, Q. Yang, J. Shan, T. Lin, J. Beyer, H. Xia, and H. Pfister. iBall: Augmenting Basketball Videos with Gaze-Moderated Embedded Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2023. 1, 2, 3, 4, 7, 9

[67] C. Zhu-Tian, Q. Yang, X. Xie, J. Beyer, H. Xia, Y. Wu, and H. Pfister. Sporthesia: Augmenting sports videos using natural language. *arXiv e-prints*, pp. arXiv–2209, 2022. 1, 3

[68] C. Zhu-Tian, S. Ye, X. Chu, H. Xia, H. Zhang, H. Qu, and Y. Wu. Augmenting Sports Videos with VisCommentator. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 2021. 1, 2, 3, 5