Sporthesia: Augmenting Sports Videos Using Natural Language

Chen Zhu-Tian, Qisen Yang, Xiao Xie, Johanna Beyer, Haijun Xia, Yingcai Wu, and Hanspeter Pfister

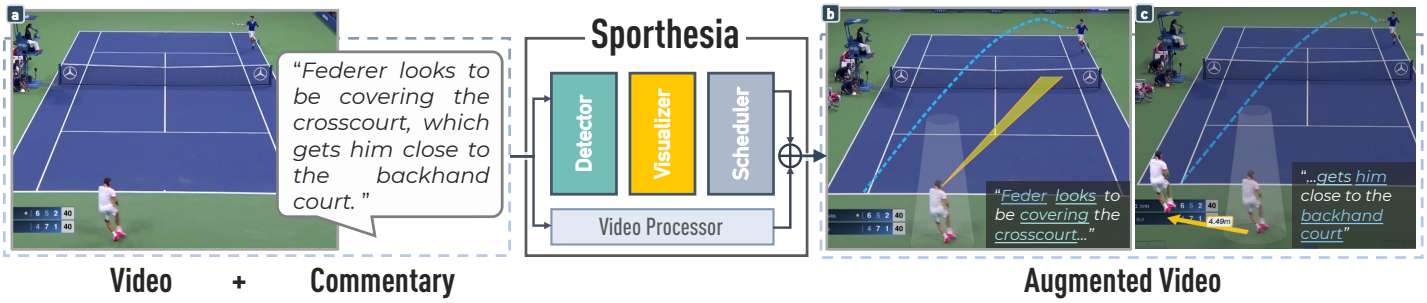

Fig. 1: Sporthesia takes raw video footage and commentary text of racket-based sports as input, and outputs an augmented video. To achieve this, three key steps are taken: 1) detecting the visualizable entities in the text, 2) mapping the entities to visualizations, and 3) scheduling the visualizations to play with the raw video.

Abstract—Augmented sports videos, which combine visualizations and video effects to present data in actual scenes, can communicate insights engagingly and thus have been increasingly popular for sports enthusiasts around the world. Yet, creating augmented sports videos remains a challenging task, requiring considerable time and video editing skills. On the other hand, sports insights are often communicated using natural language, such as in commentaries, oral presentations, and articles, but usually lack visual cues. Thus, this work aims to facilitate the creation of augmented sports videos by enabling analysts to directly create visualizations embedded in videos using insights expressed in natural language. To achieve this goal, we propose a three-step approach – 1) detecting visualizable entities in the text, 2) mapping these entities into visualizations, and 3) scheduling these visualizations to play with the video – and analyzed 155 sports video clips and the accompanying commentaries for accomplishing these steps. Informed by our analysis, we have designed and implemented Sporthesia, a proof-of-concept system that takes racket-based sports videos and textual commentaries as the input and outputs augmented videos. We demonstrate Sporthesia's applicability in two exemplar scenarios, *i.e.*, authoring augmented sports videos using text and augmenting historical sports videos based on auditory comments. A technical evaluation shows that Sporthesia achieves high accuracy (F1-score of 0.9) in detecting visualizable entities in the text. An expert evaluation with eight sports analysts suggests high utility, effectiveness, and satisfaction with our language-driven authoring method and provides insights for future improvement and opportunities.

Index Terms—Augmented Sports Videos, Language-driven Authoring Tool, Video-based Visualization, Sports Visualization

1 Introduction

Augmented sports videos are becoming increasingly popular as a form to present sports data. In recent years, an increasing amount of data has been collected during sports activities thanks to the advances in high-speed cameras and computer vision (CV) techniques. While this data plays a central role in understanding players' performance and developing winning strategies, it can be challenging to understand the data without presenting it in its physical context. Augmented sports videos can present sports data directly in actual scenes through embedded visualizations and video effects, communicating insights and explaining player strategies in an intuitive and engaging manner. Thus, augmented sports videos have been widely used by TV channels [16], fan clubs [57], and analysts [38, 51] to present sports data, influencing billions of sports enthusiasts around the world.

However, creating augmented sports videos is a demanding and time-consuming task [78], as it requires skills in areas such as data analysis, data visualizations, and video editing. The gap between the difficulty in creating augmented sports videos and the strong market demand for augmented sports videos has spawned very successful commercial systems, such as Viz Libero [58] and Piero [5]. These systems, however, target expert video editors and require the manipulation of low-level graphical elements. This leads to a high entry barrier for sports analysts, who usually focus on presenting analytical insights and lack sufficient video editing skills. Recently, Zhu-Tian et al. [78] presented VisCommentator, a fast prototyping tool for augmenting table tennis videos, enabling analysts to augment the video by interacting with the data of video objects instead of low-level graphical elements. However, analysts often express their findings as high-level insights, such as *"Federer hits a backhand down the line"*, rather than data (*e.g.*, the player's position and the ball trajectory). Consequently, to visualize an insight, an analyst usually needs to "translate" the insight into data equivalents and then map them to visualizations embedded in the videos, which is tedious and has a heavy cognitive load.

However, the tedious "translation" process implies a latent mapping between the high-level insights and the visualizations. Such a latent mapping provides an opportunity to facilitate the creation of augmented sports videos by directly creating visualizations to augment the videos based on the user's insights. In sports, perhaps the most common way to express insights is using natural language, *e.g.*, commentators give real-time comments on live games, analysts report their analytical findings of sports videos in oral presentations, and journalists summarize key events in textual documents. This work, thus, explores how to leverage natural language to facilitate the creation of augmented sports videos.

This work aims to augment a sports video clip based on a given commentary text. To achieve this goal, we first identify three tasks inspired by existing text-to-visuals systems [13, 69, 70]: 1) detecting

- *Chen Zhu-Tian, Johanna Beyer, and Hanspeter Pfister are with with John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA. E-mail: {ztchen, jbeyer, pfister}@g.harvard.edu.*
- *Qisen Yang, Xiao Xie, and Yingcai Wu are with Zhejiang University. E-mail: {qs_yang, xxie, ycwu}@zju.edu.cn. This work was done when Qisen Yang was an intern at Harvard University.*
- *Haijun Xia is with UC San Diego. E-mail: haijunxia@ucsd.edu.*

the visualizable entities in the text, 2) mapping these entities into visualizations, 3) scheduling these visualizations to play with the sports video. To tackle these three tasks, we collected and analyzed 155 sports video clips, as well as their corresponding commentaries, of six different sports with three main questions in mind – *What text entities can be visualized (Q1)?*; *How can we visualize these entities (Q2)?*; and *How do we schedule these visualizations with the video (Q3)?* Based on our formative study, we have identified that four categories of entities, namely, `object`, `action`, `data`, and `emotional cue`, in the commentaries can be visualized, entities in different categories can be visualized by different embedded visualizations, and the scheduling of visualizations depends on the style of the commentaries, *i.e.*, *analyst* or *play-by-play*, in which the video is paused or not, respectively.

Based on the findings in our formative study, we have designed and developed Sporthesia (Fig. 1), a proof-of-concept system that takes textual commentaries, racket-based sports (*e.g.*, table tennis, tennis) videos, and sports data (*e.g.*, player and ball positions, key events) as the input to produce augmented sports videos. The inputted sports data can be extracted from the videos by using CV models or manually prepared. In contrast to most existing language-driven visualization creation tools, Sporthesia sees the text as an information source instead of a command (*e.g.*, *"show me the bar chart!"*) to create the visualizations. Sporthesia features three components, *i.e.*, *Entity Detector*, *Entity Visualizer*, and *Visualization Scheduler*, to complete the aforementioned three tasks. Behind these three components is a set of state-of-the-art natural language processing (NLP) models.

Sporthesia is a technique that can be applied in different application scenarios. To demonstrate the usage of Sporthesia, we exemplify two application scenarios, including authoring augmented racket-based sports videos using text and augmenting historical sports videos based on auditory comments. To evaluate the effectiveness of Sporthesia, we first conducted a technical evaluation that focuses on the accuracy of the Entity Detector, which is the foremost step in the pipeline, and achieved a good performance of an F1 score of 0.9. A task-based expert evaluation with eight sports analysts confirmed the utility, effectiveness, and high satisfaction of our language-driven creation method. We also discuss promising future opportunities implied by observations and feedback from the study.

In summary, our main contributions are threefold: First, we conduct a formative study and subsequent identification and discussion of design considerations for augmenting sports videos based on commentary text. Second, we present the design and implementation of Sporthesia, a proof-of-concept system that augments racket-based sports videos based on a piece of commentary text. Third, we demonstrate two exemplar applications based on Sporthesia and report on a technical evaluation and expert feedback.

2 Related Work

Embedded Visualization in Sports Videos. The advances in high-speed cameras and CV techniques have made sports data from videos increasingly available [38]. To present the data in a meaningful context, researchers have proposed methods that visualize the data together with the videos, such as side-by-side [22] and embedded views [51]. This work focuses on embedded visualizations in sports videos.

The idea of embedding sports data in its physical context is not new. In the early stage, researchers embedded visualizations in court diagrams, which can be seen as a simplification of the real-world scene [3,4,37,39,60]. Recently, with more powerful graphics cards and advanced CV techniques, researchers have started to explore ways to embed visualizations in sports videos directly [21]. For example, Stein et al. [51] introduced a system that takes raw footage of soccer games as the input and automatically visualizes relevant analytic measures of the players in the video. Stein et al. [50] further extended their work with a framework that semi-automatically decides what measures should be presented at a specific moment. However, most of these research systems were developed for exploration purposes and thus do not allow users to visualize their insights in the videos for communication purposes. On the other hand, industry companies have developed commercial systems to assist in creating embedded visualizations in sports videos. For instance, Piero [5] and Viz Libero [58] are powerful video editing tools that have been used for annotating sportscasting and producing TV programs. CourtVision [11], developed by Second Spectrum [46], is a basketball watching system that automatically embeds players' status information in basketball videos to engage the audience. However, these industrial products target proficient video editors, who focus on the manipulations of graphical marks, leading to complex interface actions and a steep learning curve for sports analysts.

Perhaps the most closely related work to the present research is VisCommentator [78], which is an application that enables sports analysts to create augmented sports videos by selecting *sports data*. Our work aims to further ease the creation process and support users to augment sports videos by directly expressing *sports insights in natural language*. Compared to VisCommentator, our system is a modular technique that provides a higher abstraction level, leading to a different system design, interactions, and technical implementations.

Natural Language Interfaces for Visualization. Recent achievements in NLP have reignited interest in using natural language interfaces (NLIs) for creating data visualizations. Compared to traditional visualization creation tools, systems with NLIs enable users to express their intentions via natural language rather than interface actions or coding, thereby lowering the barrier to visualizing data.

Existing NLI systems can be roughly divided into *explicit* and *implicit* approaches. Explicit NLI systems [23,35,40,49] treat the natural language as commands and require users to illustrate their intentions explicitly. For example, DataTone [23] allows users to create visualizations of their desired data by typing, *e.g.*, *"Show me medals for hockey and skating by country"*. The NL4DV toolkit [35] takes a tabular dataset and a text query as the input and outputs visualizations in the form of JSON specifications. On the other hand, implicit NLI systems view the text descriptions as another representation of the visual content, automatically converting the text to visual content and thus enabling users to create visual content implicitly. Extensive research in computer vision, computer graphics, and human-computer interaction has explored the automatic conversion of descriptive text into visual content, such as images [69, 70], 3D shapes [9] and scenes [8, 12], documents [77], and short video clips [32]. In recent years, with the development of generative adversarial networks, a plethora of systems [29,66,71,73,74] have been proposed to generate visual content based on text descriptions. However, none of those works investigates generating augmented sports videos from textual commentaries.

Our work employs an implicit approach and enables users to create augmented videos by expressing insights in natural language. In the visualization community, only few implicit NLI systems for visualization creation exist. A representative example is Text-to-Viz [13], which extracts semantic information from the user's description of insights and maps it into static infographics. We share a similar spirit but focus on creating visualizations from text to augment sports videos, which poses extra challenges as both the text and video need to be considered in the creation process and the resulting visualizations need to be dynamic.

Auto-Generation Techniques for Data Visualization. Creating data visualizations usually involves exploring data to discover insights and mapping them into proper visualizations, both of which are demanding and time-consuming. Thus, to ease the creation process, researchers have developed techniques to automate or semi-automate the data exploration and visual mapping steps. For example, to facilitate the data exploration step, DataShot [61] employs an auto-insight technique to suggest interesting data patterns from spreadsheets and generated factsheets. By using a pattern detection engine, DataToon [27], an authoring tool for data comics, automatically suggested salient patterns of the input network data. To ease the visual mapping step, prior research has proposed template-based methods to automatically map a subset of data to different types of visualizations, including charts [65], infographics [61], and data animations [2]. The templates in these systems are usually hand-crafted based on prior knowledge and empirical studies. Research systems, such as DeepEye [31], Draco [33], VizML [26], and AutoTimeline [76], have used machine learning models to automatically learn and extract templates from existing visualizations.

In contrast to these systems, our work focuses on extracting information from natural language rather than structured datasets and converting it to embedded visualizations for augmenting sports videos. In this sense, our system is closer to Text-to-Viz [13], which generates infographics based on text, but has a very different output, augmented sports videos, which is more difficult and constrained.

Immersive Sports Visualization. Fundamentally, augmented sports videos share a similar spirit with immersive sports visualization, which leverages virtual or augmented reality devices to visualize sports data in either simulated [10, 43, 52, 56, 72] or real courts [30]. While the immersive visualizations in these systems have proven to be effective for sports data analysis, most of them are predefined and lack customizability. Thus, users cannot flexibly create new immersive visualizations to express their intentions. Our work adds to the direction of immersive sports visualization but focuses on helping users create visualizations to present data in the physical context shown in a video.

3 Text to Augmented Videos: A Three-Step Approach

The goal of this work is to augment a sports video clip based on a given commentary text by automatically converting the text into embedded visualizations. To achieve this goal, inspired by existing text-to-visuals systems [13, 69, 70], we propose a three-step approach (Fig. 2) that decomposes the problem into three tasks: 1) detecting the visualizable entities in the text, 2) mapping the entities to visualizations, and 3) scheduling the visualizations to play with the video. These tasks lead us to the following three questions: Q1–*What text entities can be visualized?*; Q2–*How can we visualize these entities?*; Q3–*How do we schedule these visualizations with the video?* To understand these questions, we conducted a formative study by collecting and analyzing 155 sports video clips and their accompanying commentaries.

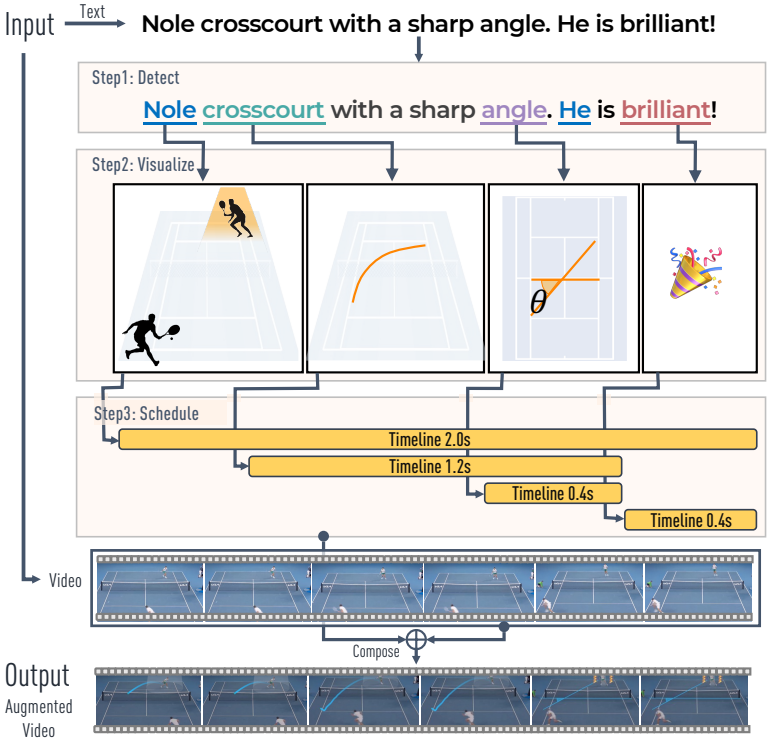

Fig. 2: A three-step approach to augment sports videos with embedded visualizations based on text commentary. The three steps include detecting visualizable entities in the text, mapping them to visualizations, and scheduling the visualizations in the video.

3.1 Data Collection and Analysis

There are many publicly available text sources that comment on sports videos, including sports commentaries, game-related reports, articles, and open discussions (*e.g.*, posts in online forums). In this work, we decided to collect sports commentaries since they are usually given by sports experts and contain rich insights.

Collection. Following the methodology in [78], we harvested a collection of commentaries that cover three team-based sports (*i.e.*, basketball, soccer, and American football) and three racket-based sports (*i.e.*, tennis, badminton, and table tennis) from the internet. Specifically, we searched sports videos with English audio commentaries from Google Videos by using the keywords "SPORT + full match", where SPORT is one of the six ball sports. We downloaded five videos for each sport from the top query results, totally gathering 30 sports videos lasting over 3600 minutes. Our collection considered both the quality (*i.e.*, millions of views) and diversity (*i.e.*, from various TV channels, including member-only ones such as ESPN+). Most of the videos were final games of famous sports events, such as the Olympic Games, FIFA world cups, NBA games, Grand Slams, and the Super Bowl. Note that our unit of analysis was not an entire game but a specific meaningful moment in the game (*e.g.*, a goal, a rally). Thus, for each video, we manually sampled at least four clips following two criteria: 1) the clip should cover a highlighted moment (curated by TV channels) of the game; 2) the commentaries of this clip should be closely related to the sports event happening in the clip (*i.e.*, commentaries about player anecdotes were thus excluded). Finally, our dataset included 155 clips lasting 92 minutes, which aligns with similar visualization research [1, 44, 54]. All videos were transcribed manually by native English speakers, resulting in 12545 words. Figure 3a and b show the average duration of the videos and the average number of words in the commentaries of different sports.

Analysis. We analyzed the dataset both qualitatively and quantitatively. Three of the authors first followed an open coding process to analyze the 155 clips independently, with the three questions (Q1-3) in mind. The codes were then refined through multiple rounds of discussions with other co-authors. The investigation also referred to prior research on text-driven visual content generation [13, 70], sports visualizations [38], data-driven videos and animated graphics [1, 44, 54], and augmented sports videos [78]. Findings revealed that there are four categories of entities that can be visualized in the text (Q1), entities in different categories can be visualized with different embedded visualizations (Q2), and these visualizations can be scheduled with the video in two different ways (Q3). We further conducted a quantitative analysis to count the occurrences of the four categories of entities in the dataset (Fig. 3c). The findings are detailed in the following sections.

3.2 Q1: What Text Entities Can Be Visualized?

Different from traditional visualization systems that generate visualizations based on structural data (*e.g.*, spreadsheets), our goal is to generate visualizations from sports commentaries, which can be unstructured, messy, and full of uncertainty. Thus, the first step to achieving our goal is to recognize the visualizable entities in the text. According to our analysis of the dataset, we identified four categories of visualizable entities (highlighted in `typewriter` font), which are introduced from concrete, objective to abstract, subjective:

- **E1** **`Objects`** are physical entities that can be seen in the video, which usually serve as the referent for other visualizations. Among all `objects`, we found two kinds of `objects` were mentioned most frequently, namely, *players* (68.83%) and *places* in the court (10.73%). Players were usually mentioned by their names and pronouns. All `objects` are noun words in the commentaries.
- **E2** **`Actions`** are performed by `objects` and can also be seen in the video. In the dataset, we found that `actions` can be roughly divided into *sports-general* (80.27%) and *sports-specific* (19.63%). Sports-general `actions`, such as *hit*, *run*, and *cover*, are usually verbs and exist in all the six ball sports. In contrast, sports-specific `actions` are terms used in specific sports, such as *down the line* in tennis, *pick and roll* in basketball. Domain-specific `actions` can be nouns (*e.g.*, *crosscourt*) or adjectives (*e.g.*, *backhand*), both of which can be used as verbs in the commentary, such as *"Federer backhand on the run," "Djokovic down the line."*

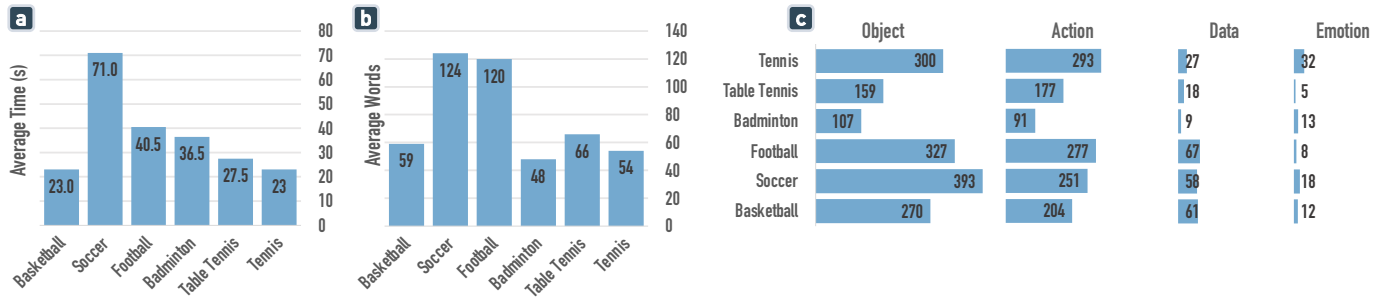

Fig. 3: The average a) duration of videos and b) number of words of commentaries per sport in the collected dataset. c) The number of entities per category in different sports.

E3 `Data` is usually generated by `actions` and cannot be seen in the video. Prior research [38] categorized `data` in sports videos into *tracking data*, which is inherently spatial and temporal, and *non-tracking data*, which is rather abstract. We also found these two kinds of data were brought up in the commentaries (16% and 84% for tracking and non-tracking data, respectively). `Data` in the commentaries usually are noun words or numbers, such as *speed*, *5 meters*, *winning rate*, and *won 64%*.

E4 `Emotional Cues` are the subjective feeling of a game expressed by the commentators, adding to the exciting atmosphere of the game. `Emotional cues` cannot be seen in the video but can be felt in the game and the commentaries. In the dataset, `emotional cues` can be adjectives (*e.g.*, *"Phenomenal!"*), interjections (*e.g.*, *"Wow!"*), or analogies (*e.g.*, *"make it just like Quidditch."*)

Relationships Among Entities. In a commentary, the visualizable entities are usually not isolated – instead, they are inherently connected. As discussed above, an `object` can perform an `action` that generates `data`. We noticed that such kinds of relationships are embedded in the linguistic structures of the natural language, which can, and should, be extracted and leveraged to specify the visualizations. For example, by identifying the subject (*i.e.*, an `object`) of an `action`, we can visualize the generated tracking data in the correct position in the video.

3.3 Q2: How to Visualize the Entities in the Video?

After recognizing the visualizable entities in the text, the next step is to map them into proper visualizations that can be embedded in the video. While prior work [78] studied the visualizations of sports data (*i.e.*, tracking or non-tracking) in videos, how to visualize commentary entities in videos remains unclear. Based on our analysis of the commentaries and the videos, multiple rounds of discussion and iterations with a domain expert (a sports science professor who provides data analysis and consultancy services for national sports teams), and the previous research on video-based sports visualizations, we propose the following methods to visualize each category of entities:

- `Object` entities can be directly seen in the video. Thus, to visualize `object` entities, we can directly *highlight* the corresponding objects in the video. Various types of `objects` can be highlighted differently. For example, we can use spotlights and rectangle marks to highlight players and places on the court, respectively.
- `Action` entities can also be seen in the video, but they do not have a physical existence. On the other hand, `actions` are performed by `objects` and generate `data`. Hence, to visualize an `action` entity, we can *highlight* the `object` when she/he/it is performing the `action` or *visualize* the `data` generated by the `action`. Compared to highlighting the subject of an `action`, visualizing the invisible `data` of the `action` can reveal more insights and better engage the audience. However, one must know what `data` is generated by the `action` to achieve the visualization, which is relatively easy for sports-general `actions` but can be challenging for sports-specific ones.
- `Data` entities, according to prior research [78], can be visualized in the video by mapping them to different visual representations based on their types. Specifically, we can naturally *embed* tracking data into the video as it is always associated with a specific space and time in the video. For non-tracking data, we can *annotate* the video with labels to show it.
- `Emotional Cues` are the most abstract entities and are sometimes not directly associated with the objects in the video. To visualize `emotional cue` entities, a straightforward way is to *display* semantic-related pictograms, such as emojis, in the video. While other more advanced methods are possible, we consider exploring them as beyond the scope of this work, as the research of affective visualization is still in its infancy [28].

In summary, entities in different categories can be visualized differently, such as highlighting the entities in the video (`object`, `action`), visualizing the data generated by the entities (`action`), embedding or annotating the entities into the video (`data`), or presenting the entities using emojis (`emotional cue`). All these methods map the entities into visualizations embedded in the video. To select the specific visualizations, we follow previous work [78] that summarized a design space of embedded visualizations in augmented sports videos.

3.4 Q3: How to Schedule the Visualizations in the Video?

After mapping the entities into visualizations, the last step is to schedule the visualizations in the video. This step entails deciding *when* to display a visualization and *how long* to display it for. This, intuitively, depends on several factors, including the text, the video, and the visualizations themselves. In our analysis, we noticed that this question is particularly related to the commentary style of the text. Specifically, there are two major types of sports commentators – *play-by-play* commentators, who need to articulate each play and event of an often fast-moving sports game, and *analyst* commentators, who provide expert analysis and background information of the game. These two types of commentaries lead to two different rendering methods, wherein the visualizations are scheduled differently:

- **Play-by-play** mode renders the visualizations without pausing the video, since the visualizations are generated by commentaries that describe the ongoing content of the video. In this mode, the scheduling of the visualizations depends on both the text and the video. For example, the visualization of an `action` should be displayed when it is mentioned in the text and disappear when the `action` is finished in the video.
- **Analyst** mode usually renders the visualizations by pausing the video, since the commentaries are often given during a break or replay of the game and contain too much information to be visualized in a short moment. In this mode, the scheduling of the visualizations only depends on the text since the video is paused. Thus, the start time and duration of a visualization should be decided by when and how it is mentioned in the text.

While we divide the rendering and scheduling methods of visualizations into two types based on the commentary styles, they are not meant to be the only solution. For example, an analyst-style commentary may also be able to be rendered without pausing the video. We leave the comprehensive exploration of the rendering and scheduling of visualizations in augmented sports videos for future research.

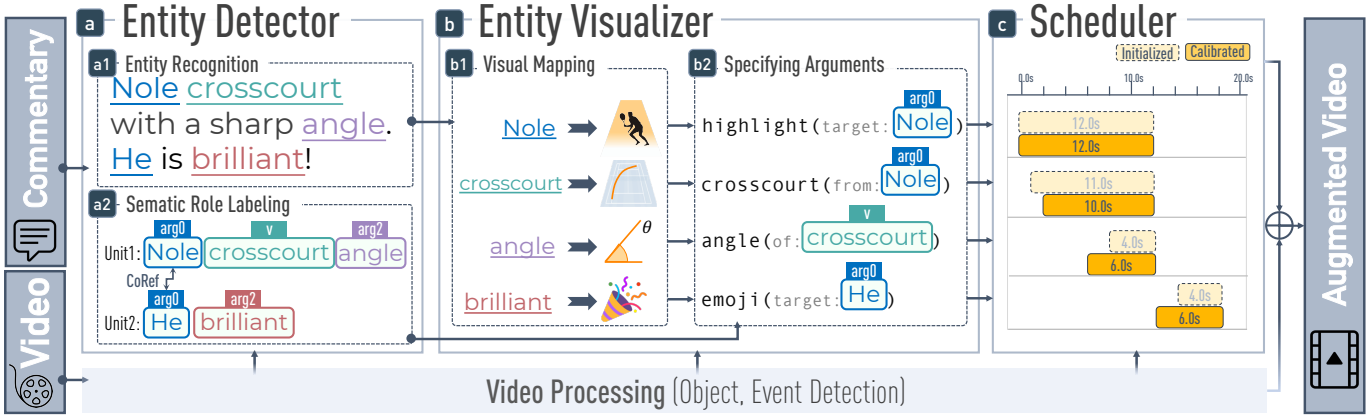

Fig. 4: Sporthesia detects the visualizable entities in the text (a1) and groups them into semantic units (a2). Next, the entities are mapped to visualizations (b1) with arguments specified by the semantic units (b2). Finally, the system initializes and calibrates the schedules of the visualizations based on the reading time of the text and the video events (c). All three steps are built upon the video processing components.

4 Sporthesia: System Design and Implementation

To realize the three-step approach, we design and implement *Sporthesia*, a proof-of-concept system that creates augmented videos for racket-based sports. Sporthesia consists of three major components – *Entity Detector*, *Entity Visualizer*, and *Visualization Scheduler* – for each step in the approach, respectively. It takes a piece of text, a video clip, and sports data extracted from the video clip as the input, and outputs an augmented video. Figure 4 displays the pipeline of Sporthesia. We first introduce the video processing and rendering techniques Sporthesia is built upon, followed by details of the three components.

4.1 Video Data Extraction and Rendering

While this research particularly focuses on leveraging natural language to create augmented sports videos, the implementation of Sporthesia is built based on advanced CV techniques. We follow previous work [78] and use several machine learning models to extract data from the videos. Specifically, we use Detectron2 [68], a detection and segmentation platform that features a mask-rcnn [24] with an ImageNet [14] pre-trained RestNet-50 [25] as the backbone to detect the bounding boxes of the players and ball, the skeletons of the players, and the court lines. We also use TTNet [59] to detect ball events, including ball bounce and net hit. For player events, we utilize the distance between the ball and the player's right hand to detect stroke events. The extracted data is used in the following steps for creating visualizations to augment the videos. We refer readers to [78] for the technical details.

To render visualizations in the videos, we further need to know the camera parameters and the player identities (*i.e.*, who is far from or near the camera). While the camera parameters can be obtained using camera calibration techniques [20, 75] on the court lines detected from the video, we treat them and the player identities as known meta information in the current implementation.

4.2 Entity Detector

The first step of the framework is to detect the visualizable entities in the text (Fig. 4a1), *i.e.*, `objects`, `actions`, `data`, and `emotional cues`. Additionally, we need to extract their relationships and group them into semantic meaningful units (Fig. 4a2). To this end, we leverage a series of state-of-the-art NLP techniques to process the input text:

Detecting Visualizable Entities. Detecting entities in a piece of text is a fundamental task in NLP called Name Entity Recognition (NER) [34], which locates and classifies segments in the text into predefined categories, such as person, location, organization, *etc*. To detect the entities in a sentence, three steps are taken: 1) tokenizing the sentence into a word sequence, 2) converting the tokens into feature vectors, and 3) classifying the feature vector of each token into categories. We achieved these three steps by using Spacy [17], an industrial-strength NLP toolkit that provides pre-trained transformer-based [15] language models to tokenize, featurize, and recognize entities in a sentence. To improve the recognition performance of Spacy, we fine-tuned its pre-trained model with the commentary examples we collected and extended its pattern matching step with sports glossaries collected from Wikipedia [62, 63].

We found that some expressions, such as *he* and *here*, refer to other entities in the text and need to be resolved to be able to visualize them. Thus, to solve this issue, we employed a neural co-reference resolution model [19], which can find all expressions that refer to the same entity in a text. In sum, the output of this step is a list of entities with their categories and pointers to their referents if they exist.

Grouping the Entities into Semantic Units. As discussed in Sec. 3.2, the entities in the text are usually connected together at the semantic level, *e.g.*, an `object` performs an `action` that generates `data`. Such semantic relationships are critical to specify the visualizations. Take *"Federer hits the ball to the backhand court"* as an example, in which four entities are detected, *i.e.*, `Federer`, `hits`, `ball`, and `backhand court`. We cannot visualize `hits` and `backhand court` without knowing the subject and possessive noun. To extract the semantic relationships among the entities, we leverage the Semantic Role Labeling (SRL) [36] technique, which detects the latent predicate-argument structure of a sentence and classifies the roles of each word. For example, `hits` will be detected as a predicate with `Federer`, `ball`, and `backhand court` as argument 0, 1, and 2, respectively. With this structure, we can easily group the entities into units like *who* (argument 0) *did what* (predicate) to *whom* (argument 1) in *what ways* (argument 2). In our implementation, we use a BERT-based model [42] to detect and classify the semantic roles of the entities. The output of this step is a list of units organized in the predicate-argument structure.

4.3 Entity Visualizer

The next step is to visualize the detected entities, which we achieve by mapping the entities to visualizations (Fig. 4b1) and specifying the visualizations' arguments based on the text (Fig. 4b2):

Mapping the Entities to Visualizations. As discussed in Sec. 3.3, different categories of entities can be mapped to different visualizations. To achieve the visual mappings, we developed a dictionary based on a design space of augmented sports videos [78] and an emoji searching engine. For example, players (`object`) are mapped to spotlight highlight effects; ball angles (`data`) are mapped to embedded visualizations in the court; *"brilliant"* (`emotional cue`) is mapped to a celebration emoji. More details can be found in the supplemental material.

A particular challenge is to map `actions` to visualizations, especially for the sports-specific `actions`, such as *crosscourt*, which are abstract and usually require case-by-case designs. Two steps are thus taken to tackle this challenge:

1. Mapping sports-general `actions` to tracking-data: In the dictionary, we manually specify the mappings from sports-general `actions` to the tracking data they generate, *i.e.*, `run` and `hit` are

mapped to the `player trajectory` and `ball trajectory`, respectively. The mappings are initialized with 21 sports-general `actions` and then extended with their synonyms using word embedding [17], *e.g.*, `hit` is extended with `stroke`, `shoot`, *etc*.

2. Mapping sports-specific `actions` to sports-general ones:
Sports-specific `actions` are terminologies that are difficult to be extended from other `actions` using synonyms matching. To map sports-specific `actions` to the tracking data in a generalizable way, we leverage sports glossaries that explain these terminologies in plain text. For instance, `crosscourt` is explained as *"Hitting the ball into the diagonal court"* in a tennis glossary [64]. Based on this explanation, we can use synonym matching to map `crosscourt` to `hit` and then map to `ball trajectory` in our dictionary.

With these two steps, we can map the `actions` to the tracking data they generated, which will be mapped to embedded visualizations.

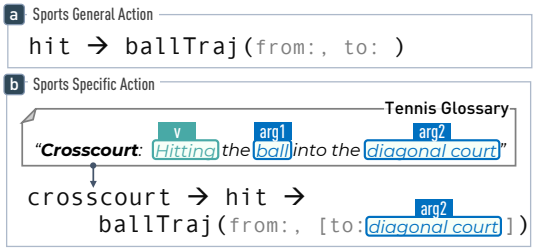

Fig. 5: a) The visualization of `hit` is manually specified, which takes two arguments, *i.e.*, *from* and *to*. b) The visualization of `crosscourt` can be generated based on its text explanation in the tennis glossary, which is a variant of `hit` with a default argument, `diagonal court`.

Specifying the Arguments of the Visualizations. To visualize the entities in the video, we further need to specify the arguments of the visualizations, which can be extracted from the text based on the semantic relationships among entities. Take *"Federer hits the ball to the backhand court"* as an example. The visualization of the `hit` action needs two arguments, *i.e.*, *from* and *to* (Fig. 5a), which can be the argument 0 (*e.g.*, `Federer`) and argument 2 (*e.g.*, `backhand court`) in the text. Some sports-specific `actions` can infer some arguments for their visualizations based on the text explanation. For example, by applying the SRL technique to the text explanation of `crosscourt` (Fig. 5b), we can set the `diagonal court` as the default value for the argument *to*. In this sense, the visualization of `crosscourt` can be seen as the one of `hit` with a default argument, `diagonal court`. More details of visualization arguments are in the supplemental material.

4.4 Visualization Scheduler

Lastly, to embed the visualizations into the video, we need to decide when a visualization should appear and disappear in the video. We initialize the time schedules of the visualizations based on the text (Fig. 6 left). The initialized schedules can be used to render the visualizations in analyst mode. When rendering in play-by-play mode, we further calibrate the schedules based on the video events (Fig. 6 right).

Initialization. Based on our analysis in Sec. 3.4, we use the text to initialize the schedules of the visualizations. Intuitively, a visualization should be displayed when its corresponding entity is read in the sentence. Thus, we employed a text-to-speech neural network [41] to generate natural speech audio for the input sentence. With the speech audio, we can obtain the start reading time of each entity, which is used as the appearance time for the visualization of the entity. Next, visualizations within the same semantic unit are scheduled to disappear at the end of reading the last word of the unit to emphasize their connections. In analyst mode (*i.e.*, when visualizations are shown as an inserted animation while pausing the video), this initial scheduling is sufficient. However, when rendering in play-by-play mode, the scheduling needs to be further calibrated.

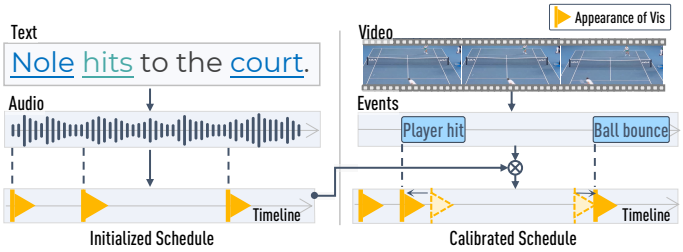

Fig. 6: Left: The text is converted into audio that initializes the appearance time of each visualization. This initialized schedule can be used to render the visualizations in analyst mode. Right: When rendering in play-by-play mode, the appearance times of some visualizations are further calibrated based on video events.

Calibration. When rendering in play-by-play mode, the visualizations of `action` and its argument entities must match the corresponding events. For example, the visualization of `hit` and `court` (Fig. 6 left) should only appear when the player hits the ball and the ball touches the ground, respectively. Thus, we calibrate the schedules of `actions` and their arguments based on the events detected in the video (Fig. 6 right). Specifically, for each `action` and its argument entities, we look up the corresponding event in the video by examining the type and time interval of the video events. If a corresponding event is found, we use the start time of the event as the appearance time for the visualization. Nevertheless, the schedules could still be sub-optimal, which should be manually refined by the users via external validity.

4.5 External Validity

As an intelligent system, Sporthesia inevitably might derive sub-optimal results due to the imperfect underlying machine learning models and the limited mapping dictionary. To support error recovery and visualization personalization, several methods for external validity can be introduced to the system. First, we can leverage the text entities as an representation to allow users to modify the system outputs. Specifically, the users can select an entity and open a context menu to modify its corresponded visualization, as well as the arguments and time schedule of the visualization. Second, the dictionary of Entity Visualizer should be configurable so that users can modify the mappings persistently. Last, the time schedules of visualizations can be visualized along with the video timeline, enabling users to understand and adjust the schedules.

5 Application Scenarios

Sporthesia is a technique that can be employed in different scenarios to augment racket-based sports videos. In this work, we implemented two application prototypes to exemplify the usage of Sporthesia.

5.1 Application I: Authoring Augmented Sports Videos

VisCommentator [78] is an authoring tool for creating racket-based augmented sports videos. We integrated Sporthesia into VisCommentator as a sub-system to enable analysts to create augmented sports videos by directly expressing high-level insights in natural language.

To achieve the integration, we modify the user interface (UI) of VisCommentator to allow text input and connect its data extractor and renderer to Sporthesia. Specifically, when a user brushes on the timeline (Fig. 7b1), a text input field will show up on the right panel (Fig. 7c). The user can then type in the input field to comment on the video. Once the user presses the play button, the text and the data extracted from the video will be passed to Sporthesia, which generates and schedules embedded visualizations through the three components (Fig. 4). The scheduled embedded visualizations will be rendered into the video by the renderer of VisCommentator.

Additionally, we also provide UI to support external validity of Sporthesia. The text entities visualized in the video will be highlighted in the text input field (Fig. 7c). Users can right click on an entity to assess and modify how it is visualized. For example, a user can configure the visualization and schedule of the entity `distance` in Fig. 7d.

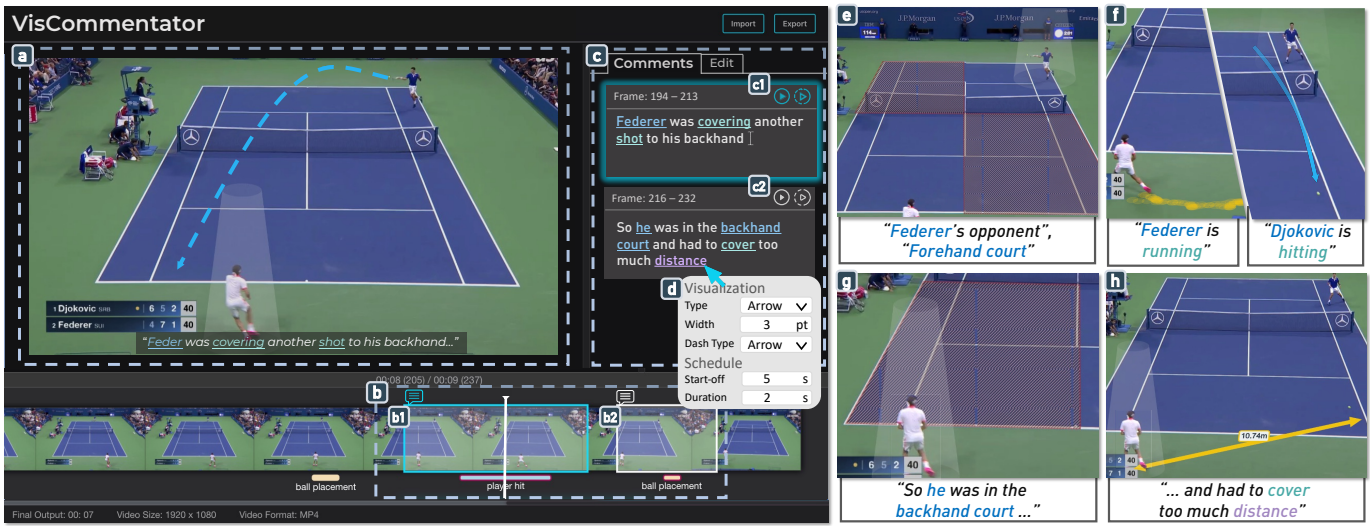

Fig. 7: The user interface of VisCommentator after being integrated with Sporthesia. Users can brush on the timeline and comment on the brushed period to e) highlight the players or places, f) visualize tracking data, and g) - h) explain insights. Users can also d) assess and modify the visualization and schedules of the text entities.

Users can also manually create or remove an entity to create their own visualizations if the detection is incorrect, and switch to the "edit" panel to modify other settings such as the visual mapping dictionary.

After integrating with Sporthesia, VisCommentator enables users to create augmented sports videos more efficiently. For example, users can simply highlight objects or actions in the video through text (Fig. 7d and e). Besides, users can comment on a moment using a more complex sentence (Fig. 1) or on multiple moments (Fig. 7b1 and b2) to create a series of augmented clips (Fig. 7a, f, and g).

5.2 Application II: Augmenting Archive Sports Videos

Many sports events, such as the Wimbledon Championships, are broadcast on TV, recorded as videos with audio commentaries, and released on online platforms (*e.g.*, YouTube). While these videos are widely used for analysis or entertainment purposes, the audio commentaries are usually not fully exploited. A promising use case is to leverage the audio commentaries to generate visualizations to augment the video, thereby facilitating the analysis of the games and increasing the engagement of watching experiences.

To augment racket-based sports videos by leveraging the audio commentaries, we implemented an examplar application that integrates a speech-to-text (STT) neural network with Sporthesia. Specifically, the application takes a video clip with audio commentaries as the input, separates the audio track from the video, converts the audio commentaries into text by using Silero [53], an enterprise-grade pre-trained STT model, and finally generates the augmented video by using Sporthesia.

Figure 8 presents an example produced by the application. The input video is a BBC sports clip of the women's final in the 2019 Wimbledon Championships. In the generated augmented video, the ball trajectory and the court inside the baseline are visualized when the commentator describes that Halep (*i.e.*, the player far from the camera) hits the ball *"just inside the baseline"* (Fig. 8a). Next, the ball trajectory is visualized in yellow color to represent *"Williams crosscourt."* (Fig. 8b) Right after Halep's forehand down the line, the commentator complimented that *"this is brilliant"*, which is visualized as several celebration emojis around Halep in the augmented video (Fig. 8c). Finally, the augmented video displays Halep's running trajectory followed by the ball trajectory and trophy emojis (Fig. 8d) when the commentator said *"She's now running onto the backhand and got a crosscourt for the winner."*

6 Evaluation

This section reports technical and expert evaluations on Sporthesia.

6.1 Technical Evaluation

We evaluated the accuracy of recognizing the four categories of text entities (Fig. 4a1), which is the foremost step of our approach. We did not technically evaluate other components because they are either off-the-shelf components with high accuracy (*e.g.*, the SRL model achieves a 0.86 F1-score on benchmark datasets) or the ground truth is not available (*e.g.*, the visualizations and schedules). Instead, we conducted expert interviews to collect qualitative feedback on the overall system.

Dataset. We prepared a dataset for entity recognition by labeling the entities in each sentence of our collected commentaries (Fig. 3c). Specifically, we manually labeled the start and end index of each entity as well as its category according to our analysis in Sec. 3. The label of each entity was represented as $(start, end, category)$. We randomly split the dataset into 10 participants for 10-fold cross-validation.

Model. To recognize the entities in a sentence, we use our dataset to fine-tune the pre-trained *en_core_web_trf* [45] model provided by Spacy [17]. The *en_core_web_trf* model is trained on written text such as blogs, news, comments using the transformer structure [15]. We also extend its pattern matching step [18] with sports glossaries. For each

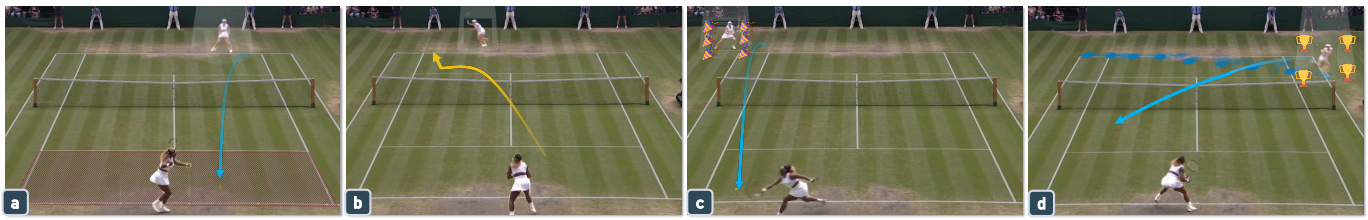

Fig. 8: Sporthesia leverages audio commentaries to augment historical sports videos with embedded visualizations.

round of cross-validation, we use nine partitions to fine-tune the model and use the remaining one for the testing.

Table 1: Precision, recall, and F1-score of the entity recognition.

Entity	Object	Action	Data	Emotion Cue
Precision	0.92	0.90	0.86	0.84
Recall	0.95	0.95	0.90	0.91
F1-Score	0.93	0.92	0.88	0.88

Results. Table. 1 shows the mean precision, recall, and F1-scores of the 10-fold cross-validation. Overall, the model achieves high accuracy across the four categories. The accuracy of `object` and `action` is comparatively higher than those of `data` and `emotional cue` since the latter have fewer data points. Note that the model, as well as its pattern matching, is data-driven, which means our Entity Detector can be improved by and generalized to other larger datasets.

6.2 Expert Evaluation

To assess the utility and effectiveness of Sporthesia, we used VisCommentator as a technology probe to conduct a qualitative expert evaluation. The study aimed to evaluate whether sports analysts can create augmented videos with our system, observe their creation process, reflect on future improvements, and collect feedback about the three-step approach and how language-driven authoring can facilitate their overall workflows of presenting analytical findings.

Participants: We recruited 8 sports analysts (P1-P8; 3 female; age: 20-58) from a university sports science department. All experts majored in Sports Training with proficient experience in analyzing racket-based sports matches. P1-P4 particularly focused on analyzing tennis matches, while P5-P8 focused on table tennis. P8 was a senior sports analyst lead who had more than ten years of experience in providing consulting services for national sports teams. All the experts only had experience with lightweight video editors, *e.g.*, Tiktok [55], rather than advanced video editing tools such as Adobe Premiere. Each participant received a gift card worth $16 at the beginning of the session, independent of their performance.

Tasks: The participants were asked to finish a training task and two creation tasks by using VisCommentator. For the first creation task, we curated a raw video **T1** that contained more than five turns, in which a player lost the rally due to an unforced error, and required the experts to augment the video using commentaries in play-by-play style. The second creation task required the experts to use analyst-style comments on a video **T2**, in which a player lost the rally due to a forced error in less than five turns. The training task was prepared to cover all the features in the two creation tasks by providing a video **T0** with more than five turns and a player lost due to a forced error. In total, we prepared six raw videos, each three for tennis and table tennis, respectively. The original commentaries of each video were removed. We also provided the experts with a document of the vocabulary and example sentences supported by the system.

Procedure: The study began with the introduction (10 min) of the study purpose, the concept of augmented sports videos, the motivation of language-driven authoring, and the concepts in our three-step approach. Next, we proceeded to the training task (15 min). We demonstrated the features of the system with video **T0** and asked the experts to reproduce the augmented videos themselves.

After the training, we provided the experts with two raw videos (**T1** and **T2**) for the two creation tasks (15min for each). We encouraged the experts to watch and ask questions about the videos before the creation. For each creation task, we asked the experts to comment on at least three segments of the video and to create eight visual effects. Finally, the session ended with a semi-structured interview (15 min) and a post-study questionnaire (5-Point Likert Scale). Each session was run in-lab using a 27-inch monitor, following a think-aloud protocol.

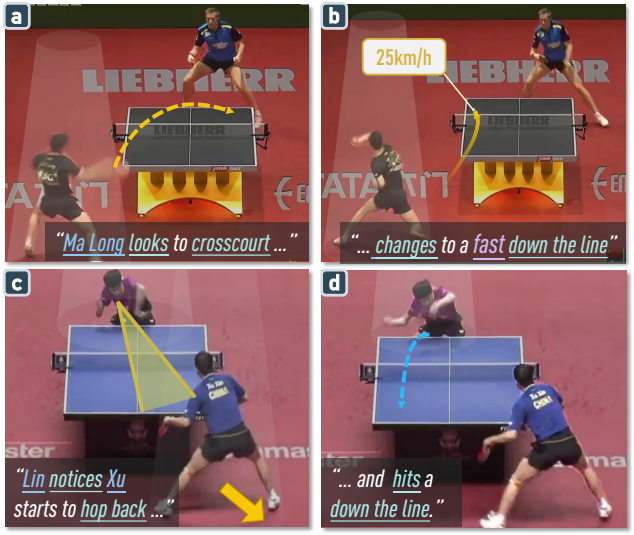

Fig. 9: Two examples, a-b and c-d, created by the experts in the study.

6.2.1 Results

All experts successfully created multiple augmented videos by using VisCommentator in the creation tasks. Figure 9 shows two examples of augmented table tennis videos created by the experts during the study. Unsurprisingly, all experts spoke highly of the usability of the system, considering it *"easy to learn and use"*. These results qualitatively demonstrated the efficiency and usability of our system. The experts' feedback is summarized as below:

Usefulness: All experts confirmed the usefulness ($\mu = 4.88$) of our language-driven authoring method for sports analysts to create augmented sports videos, as it *"significantly lowers the entry barriers to augmented sports videos for analysts"* (P4). The experts expressed that the usefulness of our method is rooted in its *"intuitiveness"* and *"efficiency"* (P1-8), which allows analysts to create augmented sports videos in a short period without being tangled in the details of video editing. Particularly, the experts lauded that our language-driven method *"occupies a unique niche"* (P2) of fast prototyping augmented videos for day-to-day usages, such as discussion, presentation, and demonstration. The experts pointed out that in these scenarios, augmented videos can significantly facilitate the communication of sports insights but are unnecessary to be high-fidelity. Thus, existing video editing tools are too "heavy" while our method *"can perfectly fill this gap"* (P8).

Effectiveness: The design of our three-step approach was rated as effective by the experts ($\mu = 4.65$). The experts thought that the four categories of entities purposed by us were *"reasonable"* (P3-8) and *"sufficient to present sports insights"* (P3, 6-8). Some experts (P1, 2, 5) suggested that the system should detect some deep semantic meanings, such as the players' emotions and the situation of the games. The experts also agreed with our proposed visualization methods for each category and suggested that the players' emotions can be visualized by highlighting their actions (P7). As for the two scheduling methods, while most of the experts appreciated our proposed designs, some experts who focused on table tennis (P7-8) thought that play-by-play mode is not that useful as in table tennis usually both the players and ball move too fast to be augmented.

Satisfaction: The rating also reflects positive user satisfaction for the implementation of Sporthesia ($\mu = 4.23$). The experts indicated that the detector could correctly recognize the key information in their comments and the scheduler could properly arrange the visualizations by incorporating both the text orders and video events. However, comments also suggested that the dictionary that maps entities to visualizations was not comprehensive enough so that *"I have to find alternative expressions to describe a tactic."* (P6) We considered this can be mitigated in future improvement by extending the dictionary.

6.2.2 Observations, Feedback, and Future Opportunities

Visualizing Deep Semantic Information in Physical Contexts. Some experts suggested that the system should be able to detect and visualize deep semantic meanings in the text. For example, P6 wanted to visualize the *"tense situation"* in the game; P1 noted that although he *"comments the same [kinds of] actions"*, the visualizations should be different since he may have *"different tendencies."* However, extracting and visualizing the deep semantic information are both challenging tasks, which requires further study in NLP and visualizations. From the perspective of visualization, visualizing such kind of highly abstract information in a physical context remains underexplored. Recently, research in Situated Visualization [7], an emerging research topic, has conducted preliminary exploration in this direction.

Collaborative Interactions Across Abstraction Levels. While natural language can allow users to convey high-level insights, we notice that sometimes the expert wants to express low-level information that can hardly be expressed in language. For example, P1 found that it was difficult to express a specific court location in language. Instead, he would like to type *"Federer moves to ..."* and then use the mouse to directly click the location in the video. This interesting observation implies that when authoring augmented videos, users need interactions with different expressiveness ranging from low to high abstraction levels. However, how to unify the interactions across various abstraction levels remains an open question. Recently, Srinivasan et al. [48] explored consistent multimodal interactions for data visualization on tablets, providing relevant knowledge in this interesting direction.

Opportunities Enabled by NLIs to Bridge Data Analysis and Communication. Surprisingly, the experts suggested that the NLI system not only facilitates the creation of augmented videos but also their analysis process. P5 provided that the augmentations generated based on comments can be seen as visual notes of the video, which helps him externalize and organize his thoughts. *"At the beginning [of analysis,] my thoughts are fragmented.."*, P5 detailed that, *"...visualizing them can help me to think."* Such feedback suggests a promising opportunity to bridge the gap between data analysis and communication. Specifically, with our technique, a visual analytics system for sports videos can enable users to take textual notes on the videos, visualize their notes to facilitate the analysis, and gradually shift to authoring augmented sports videos to present the analytical findings. We discussed this new workflow with the experts and received very positive feedback. Thus, we suggest further exploration in this direction.

Suggestions. The experts also identified some limitations, most of which were related to system engineering maturity. For example, compositions of multiple augmentations were not supported. The experts did come across certain issues rooted in the design of our language-driven method. First, we noticed that the experts were hesitant to type when guessing which words the system could understand. Such an issue is a long-standing challenge for users of NLI systems [47]. One plausible solution is to integrate auto-complete features into the system. Second, the experts showed that the visual mappings should adapt to the specific sports. For example, the visualizations of distances between a player and a place should be different in tennis and table tennis (*i.e.*, stand on the court *vs.* behind the table). Finally, the reading times of entities and the video events can be mismatched. The future system should provide interactive functions for manual corrections.

7 Discussion

Failure Cases Due to ML Models. During our study, we observed some failure cases in which Sporthesia cannot create augmented videos correctly. One major source of these cases is the imperfect NLP models, especially the SRL model. For some complex and long sentences, such as *"Fan, now with the side of the racket that makes it spring off the rubber fast"*, the SRL model cannot extract the desired predict-arguments structures, leading to problematic visualizations. Moreover, the SRL model can only extract the shallow semantic meanings of the sentence but cannot understand the deep semantic relationships among entities or sentences. For example, *"Dema takes a swatting backhand after this inside-out forehand from Zhendong"*, will lead to an error order of what commentators want to convey. In addition, the underlying CV models may also cause inappropriate rendering results due to, for example, incorrect object detection, tracking, or segmentation. Nevertheless, these issues can be addressed or mitigated with more advanced models, larger datasets, and better implementations [67].

Generalizability—beyond racket-based sports. While Sporthesia is designed and implemented for tennis and table tennis, it can be generalized to other racket- or team-based sports as it is built based on a formative study of both racket- and team-based sports. Among the three components, the bottleneck for generalization lies in the Entity Visualizer, as the other two–Entity Detector (data-driven) and Visualization Scheduler (domain-agnostic)–are naturally generalizable. Entity Visualizer needs to be extended with a domain specific dictionary that maps sports actions to embedded visualizations. Such a verb-visual dictionary will increase the expressive power of Sporthesia, but also contribute to the visualization community in many directions, such as data animations and NLI systems. Another bottleneck of extending Sporthesia to team-based sports is the underlying CV models, which require detecting and tracking multiple objects. Recent advances in transformer-based models [6] for video processing can be a solution.

Applicability—broader application scenarios. We have showcased that Sporthesia can be employed in various application scenarios. In the user study, the experts also suggested multiple interesting applications. On one hand, the experts believed that our technique could benefit presenters in scenarios such as teaching, group discussion, and TV broadcasting. The experts commented that when presenting insights about sports videos, the generated visualizations can *"reduce the ambiguity [of spoken language] and facilitate the communication"* (P8). On the other hand, some experts also indicated that our technique can be used in game-viewing systems for audiences. P8 explained that *"fans can type comments to highlight players or visualize data in game watching."* We consider leveraging interactive embedded visualizations to improve game watching experiences as an interesting future direction.

Study Limitations. Since Sporthesia was built based on findings derived from English spoken commentaries and was only implemented for tennis and table tennis and covered the words discovered from the collected commentaries. Further studies or adaptations may still be necessary when generalizing it to other scenarios. Besides, the expert evaluations only provided qualitative feedback since the sample size is small due to the limited nature of access to experts. Finally, although the experts were satisfied with the created augmented videos, we didn't evaluate the videos from the audience's perspective. Follow-up empirical evaluation in real-world settings is thus suggested.

8 Conclusion

This work aims to facilitate the creation of augmented sports videos using insights expressed in natural language. To achieve this goal, we proposed a three-step approach inspired by existing research in text-to-visuals. We conducted a formative study to analyze 155 augmented sports videos and their commentaries to answer three key questions in the approach. Informed by the analysis results, we designed and implemented Sporthesia, a proof-of-concept system that creates augmented videos for racket-based sports using textual comments. To demonstrate the applicability of Sporthesia, we presented two application scenarios, *i.e.*, authoring augmented sports videos and augmenting historical sports videos based on auditory comments. A technical evaluation showed that Sporthesia can successfully detect visualizable text entities. A user study with eight sports analysts revealed the utility, effectiveness, and high satisfaction of the system. Feedback and observations from the study suggest promising future research directions.

Acknowledgments

The authors wish to thank the sports experts from Zhejiang University for their time and expertise. A special thanks to Salma Abdel Magid for her beautiful voice and help on the video narration. This research is supported in part by the NSF award III-2107328, NSF award IIS-1901030, NIH award R01HD104969, and the Harvard Physical Sciences and Engineering Accelerator Award.

REFERENCES

[1] F. Amini, N. H. Riche, B. Lee, C. Hurter, and P. Irani. Understanding Data Videos: Looking at Narrative Visualization through the Cinematography Lens. In *Proc. of CHI*, pp. 1459–1468. ACM, 2015.

[2] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, and P. Irani. Authoring Data-Driven Videos with DataClips. *IEEE TVCG*, 23(1):501–510, 2017.

[3] G. L. Andrienko, N. V. Andrienko, G. Anzer, P. Bauer, G. Budziak, G. Fuchs, D. Hecker, H. Weber, and S. Wrobel. Constructing Spaces and Times for Tactical Analysis in Football. *IEEE TVCG*, 27(4):2280–2297, 2021. doi: 10.1109/TVCG.2019.2952129

[4] G. L. Andrienko, N. V. Andrienko, G. Budziak, T. von Landesberger, and H. Weber. Exploring pressure in football. In *Proc. of AVI*, pp. 54:1–54:3. ACM, 2018. doi: 10.1145/3206505.3206558

[5] BBC. Piero. `https://www.bbc.co.uk/rd/projects/piero`, 2022.

[6] G. Bertasius, H. Wang, and L. Torresani. Is Space-Time Attention All You Need for Video Understanding? In *Proc. of ICML*, July 2021.

[7] N. Bressa, H. Korsgaard, A. Tabard, S. Houben, and J. Vermeulen. What's the Situation with Situated Visualization? A Survey and Perspectives on Situatedness. *IEEE TVCG*, 28(1):107–117, 2022.

[8] A. X. Chang, M. Savva, and C. D. Manning. Learning Spatial Knowledge for Text to 3D Scene Generation. In *Proc. of EMNLP*, pp. 2028–2038. ACL, 2014. doi: 10.3115/v1/d14-1217

[9] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese. Text2Shape: Generating Shapes from Natural Language by Learning Joint Embeddings. In *Proc. of ACCV*, vol. 11363, pp. 100–116. Springer, 2018. doi: 10.1007/978-3-030-20893-6_7

[10] 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 TVCG*, 28(1):118–128, 2022. doi: 10.1109/TVCG.2021.3114861

[11] Clippers. Court vision. `https://www.clipperscourtvision.com/`, 2022.

[12] B. Coyne and R. Sproat. WordsEye: An Automatic Text-to-Scene Conversion System. In *Proc. of SIGGRAPH*, pp. 487–496. ACM, 2001. doi: 10.1145/383259.383316

[13] W. Cui, X. Zhang, Y. Wang, H. Huang, B. Chen, L. Fang, H. Zhang, J. G. Lou, and D. Zhang. Text-to-Viz: Automatic Generation of Infographics from Proportion-Related Natural Language Statements. *IEEE TVCG*, 26(1):906–916, 2020. doi: 10.1109/TVCG.2019.2934785

[14] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A Large-scale Hierarchical Image Database. In *Proc. CVPR*, pp. 248–255. IEEE, 2009.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of ACL*, pp. 4171–4186. ACL, June 2019. doi: 10.18653/v1/N19-1423

[16] ESPN. Detail. `https://www.espn.com/watch/catalog/f48c68af-f980-4fcb-8b59-2a0db01f50cf/_/country/us`, 2022.

[17] Explosion.Ai. SpaCy. `https://spacy.io/`, 2022.

[18] Explosion.Ai. SpaCy - Pattern Matching. `https://spacy.io/usage/rule-based-matching`, 2022.

[19] H. Face. NeuralCoref 4.0. `https://github.com/huggingface/neuralcoref`, 2022.

[20] D. Farin, S. Krabbe, P. H. N. de With, and W. Effelsberg. Robust Camera Calibration for Sport Videos Using Court Models. In *Proc. of Storage and Retrieval Methods and Applications for Multimedia*, vol. 5307, pp. 80–91. SPIE, 2004. doi: 10.1117/12.526813

[21] M. T. Fischer, D. A. Keim, and M. Stein. Video-based Analysis of Soccer Matches. In *Proc. of International Workshop on Multimedia Content Analysis in Sports*, pp. 1–9, 2019.

[22] Y. Fu and J. T. Stasko. Supporting data-driven basketball journalism through interactive visualization. In *Proc. of CHI*, pp. 598:1–598:17. ACM, 2022. doi: 10.1145/3491102.3502078

[23] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. Karahalios. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proc. of UIST*, pp. 489–500. ACM, 2015. doi: 10.1145/2807442.2807478

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proc. of ICCV*, pp. 2961–2969. IEEE, 2017.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*, pp. 770–778. IEEE, 2016.

[26] K. Z. Hu, S. N. S. Gaikwad, M. Hulsebos, M. A. Bakker, E. Zgraggen, C. A. Hidalgo, T. Kraska, G. Li, A. Satyanarayan, and Ç. Demiralp. VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository. In *Proc. of CHI*, p. 662. ACM, 2019.

[27] N. W. Kim, N. H. Riche, B. Bach, G. Xu, M. Brehmer, K. Hinckley, M. Pahud, H. Xia, M. J. McGuffin, and H. Pfister. DataToon: Drawing Dynamic Network Comics With Pen + Touch Interaction. In *Proc. of CHI*, p. 105. ACM, 2019.

[28] X. Lan, Y. Shi, Y. Wu, X. Jiao, and N. Cao. Kineticharts: Augmenting Affective Expressiveness of Charts in Data Stories with Animation Design. *IEEE TVCG*, 28(1):933–943, 2022.

[29] Y. Li, M. R. Min, D. Shen, D. E. Carlson, and L. Carin. Video Generation From Text. In *Proc. of AAAI*, pp. 7065–7072. AAAI Press, 2018.

[30] T. Lin, R. Singh, Y. Yang, C. Nobre, J. Beyer, M. A. Smith, and H. Pfister. Towards an Understanding of Situated AR Visualization for Basketball Free-Throw Training. In *Proc. of CHI*, pp. 461:1–461:13. ACM, 2021. doi: 10.1145/3411764.3445649

[31] Y. Luo, X. Qin, N. Tang, and G. Li. DeepEye: Towards Automatic Data Visualization. In *Proc. of ICDE*, pp. 101–112. IEEE, 2018.

[32] T. Marwah, G. Mittal, and V. N. Balasubramanian. Attentive Semantic Video Generation Using Captions. In *Proc. of ICCV*, pp. 1435–1443. IEEE, 2017. doi: 10.1109/ICCV.2017.159

[33] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing Visualization Design Knowledge as Constraints: Actionable and Extensible Models in Draco. *IEEE TVCG*, 25(1):438–448, 2019.

[34] D. Nadeau and S. Sekine. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[35] A. Narechania, A. Srinivasan, and J. Stasko. NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. *IEEE TVCG*, 27(2):369–379, 2021. doi: 10.1109/TVCG.2020.3030378

[36] M. Palmer, D. Gildea, and N. Xue. Semantic Role Labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103, 2010.

[37] C. Perin, R. Vuillemot, and J. Fekete. SoccerStories: A Kick-off for Visual Soccer Analysis. *IEEE TVCG*, 19(12):2506–2515, 2013. doi: 10.1109/TVCG.2013.192

[38] C. Perin, R. Vuillemot, C. D. Stolper, J. T. Stasko, J. Wood, and S. Carpendale. State of the Art of Sports Data Visualization. In *Proc. of CGF*, vol. 37, pp. 663–686. Wiley Online Library, 2018.

[39] D. Sacha, F. Al-amoody, M. Stein, T. Schreck, D. A. Keim, G. L. Andrienko, and H. Janetzko. Dynamic Visual Abstraction of Soccer Movement. *CGF*, 36(3):305–315, 2017. doi: 10.1111/cgf.13189

[40] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A Natural Language Interface for Visual Analysis. In *Proc. of UIST*, pp. 365–377. ACM, 2016. doi: 10.1145/2984511.2984588

[41] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al. Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. In *Proc. of ICASSP*, pp. 4779–4783. IEEE, 2018.

[42] P. Shi and J. Lin. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *ArXiv*, abs/1904.05255, 2019.

[43] S. Shimizu and K. Surni. Sports Training System for Visualizing Bird's-Eye View from First-Person View. In *Proc. of VR*, pp. 1156–1158. IEEE, 2019.

[44] X. Shu, A. Wu, J. Tang, B. Bach, Y. Wu, and H. Qu. What Makes a Data-GIF Understandable? *IEEE TVCG*, 27(2):1492–1502, 2021.

[45] Spacy. English transformer pipeline. `https://spacy.io/models/en_core_web_trf`, 2022.

[46] S. Spectrum. Second spectrum. `http://secondspectrum.com/`, 2022.

[47] A. Srinivasan, M. Dontcheva, E. Adar, and S. Walker. Discovering Natural Language Commands in Multimodal Interfaces. In *Proc. of IUI*. ACM, 2019. doi: 10.1145/3301275.3302292

[48] A. Srinivasan, B. Lee, N. H. Riche, S. M. Drucker, and K. Hinckley. InChorus: Designing Consistent Multimodal Interactions for Data Visualization on Tablet Devices. In *Proc. of CHI*, pp. 1–13. ACM, 2020. doi: 10.1145/3313831.3376782

[49] A. Srinivasan, N. Nyapathy, and B. Lee. Collecting and Characterizing Natural Language Uterances for Specifying Data Visualizations. In *Proc. of CHI*. ACM, 2021. doi: 10.1145/3411764.3445400

[50] M. Stein, T. Breitkreutz, J. Häussler, D. Seebacher, C. Niederberger, T. Schreck, M. Grossniklaus, D. A. Keim, and H. Janetzko. Revealing the Invisible: Visual Analytics and Explanatory Storytelling for Advanced Team Sport Analysis. In *Proc. of BDVA*, pp. 1–9. IEEE, 2018.

[51] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlücke, T. Schreck, G. Andrienko, M. Grossniklaus, and D. A. Keim. Bring It to the Pitch: Combining Video and Movement Data to Enhance Team Sport Analysis. *IEEE TVCG*, 24(1):13–22, 2017.

[52] Y. Tanaka, T. Shiokawa, and M. Shiokawa. Scope of Manipulability Sharing: A Case Study for Sports Training. In *Proc. of VR*, pp. 701–702. IEEE, 2018.

[53] S. Team. Silero Models: Pre-trained Enterprise-grade STT / TTS Models and Benchmarks. `https://github.com/snakers4/silero-models`, 2021.

[54] J. Thompson, Z. Liu, W. Li, and J. T. Stasko. Understanding the Design Space and Authoring Paradigms for Animated Data Graphics. *CGF*, 39(3):207–218, 2020.

[55] Tiktok. Tiktok. `https://www.tiktok.com`, 2022.

[56] W.-L. Tsai, T.-Y. Pan, and M.-C. Hu. Feasibility Study on Virtual reality Based Basketball Tactic Training. *IEEE TVCG*, 2020.

[57] T. TV. Tennis tv. `https://www.facebook.com/watch/TennisTV`, 2022.

[58] Vizrt. Viz libero. `https://www.vizrt.com/products/viz-libero`, 2022.

[59] R. Voeikov, N. Falaleev, and R. Baikulov. TTNet: Real-Time Temporal and Spatial video Analysis of Table Tennis. In *Proc. of CVPR*, pp. 3866–3874. IEEE, 2020.

[60] J. Wang, J. Wu, A. Cao, Z. Zhou, H. Zhang, and Y. Wu. Tac-Miner: Visual Tactic Mining for Multiple Table Tennis Matches. *IEEE TVCG*, 27(6):2770–2782, 2021. doi: 10.1109/TVCG.2021.3074576

[61] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, and D. Zhang. DataShot: Automatic Generation of Fact Sheets from Tabular Data. *IEEE TVCG*, 26(1):895–905, 2020.

[62] Wikipedia. Glossary of table tennis. `https://en.wikipedia.org/wiki/Glossary_of_table_tennis`, 2022.

[63] Wikipedia. Glossary of tennis terms. `https://en.wikipedia.org/wiki/Glossary_of_tennis_terms`, 2022.

[64] Wikipedia. Glossary of Tennis Terms, 2022.

[65] K. Wongsuphasawat, D. Moritz, A. Anand, J. D. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE TVCG*, 22(1):649–658, 2016.

[66] A. Wu and H. Qu. Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks. *IEEE transactions on visualization and computer graphics*, 26(7):2429–2442, 2018.

[67] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang, D. Zhang, and H. Qu. Ai4vis: Survey on artificial intelligence approaches for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[68] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[69] H. Xia. Crosspower: Bridging Graphics and Linguistics. In *Proc. of UIST*, pp. 722–734. ACM, 2020. doi: 10.1145/3379337.3415845

[70] H. Xia, J. Jacobs, and M. Agrawala. Crosscast: Adding Visuals to Audio Travel Podcasts. In *Proc. of UIST*, pp. 735–746. ACM, 2020. doi: 10.1145/3379337.3415882

[71] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks. In *Proc. of CVPR*, pp. 1316–1324. IEEE, 2018. doi: 10.1109/CVPR.2018.00143

[72] S. Ye, C. Zhu-Tian, X. Chu, Y. Wang, S. Fu, L. Shen, K. Zhou, and Y. Wu. ShuttleSpace: Exploring and Analyzing Movement Trajectory in Immersive Visualization. *IEEE TVCG*, 27(2):860–869, 2021.

[73] H. Zhang, T. Xu, and H. Li. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks. In *Proc. of ICCV*, pp. 5908–5916. IEEE, 2017. doi: 10.1109/ICCV.2017.629

[74] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE TPAMI*, 41(8):1947–1962, 2019. doi: 10.1109/TPAMI.2018.2856256

[75] Z. Zhang. A Flexible New Technique for Camera Calibration. *IEEE TPAMI*, 22(11):1330–1334, 2000. doi: 10.1109/34.888718

[76] C. Zhu-Tian, Y. Wang, Q. Wang, Y. Wang, and H. Qu. Towards Automated Infographic Design: Deep Learning-based Auto-Extraction of Extensible Timeline. *IEEE TVCG*, 26(1):917–926, 2020.

[77] C. Zhu-Tian and H. Xia. CrossData: Leveraging Text-Data Connections for Authoring Data Documents. In *Proc. of CHI*, pp. 95:1–95:15. ACM, 2022. doi: 10.1145/3491102.3517485

[78] C. Zhu-Tian, S. Ye, X. Chu, H. Xia, H. Zhang, H. Qu, and Y. Wu. Augmenting Sports Videos with Viscommentator. *IEEE TVCG*, 28(1):824–834, 2021.