AdaPIP: Adaptive Picture-in-Picture Guidance for 360° Film Watching

Yi-Xiao Li[1] Guan Luo[1] Yi-Ke Xu[1] Yu He[2] Fang-Lue Zhang[3] Song-Hai Zhang[1]
[1]BNRist, Tsinghua University, Beijing, China
[2]Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, China
[3]Victoria University of Wellington, Wellington, New Zealand
{liyixiao20@mails.,lg22@mails.,shz@}tsinghua.edu.cn
909333601@qq.com hooyeeevan2511@gmail.com fanglue.zhang@vuw.ac.nz

Abstract

360° videos enable viewers to watch freely from different directions but inevitably prevent them from perceiving all helpful information. Picture-In-Picture (PIP) guidance is proposed using preview windows to show regions of interest (ROIs) out of the current view range to mitigate this problem. We identify several drawbacks of this representation and propose a new method for 360° film-watching named AdaPIP. AdaPIP enhances the traditional PIP by adaptively arranging preview windows with changeable view ranges and sizes. Besides, AdaPIP incorporates the advantage of arrow-based guidance by presenting circular windows with arrows attached to them to help users find the corresponding ROIs more efficiently. We also adapt AdaPIP and Outside-In to HMD-based immersive virtual reality environments, demonstrating the usability of PIP-guided approaches beyond 2D screens. Comprehensive user experiments on 2D screens as well as in VR environments indicate that AdaPIP is superior to alternative methods in terms of visual experiences while maintaining a comparable degree of immersion.

Keywords:360° Videos; Picture-In-Picture; Virtual Reality; Visual Guidance

1. Introduction

Panoramic videos, also known as 360° videos, allow filmmakers to produce dynamic scenes that support viewers to watch from different virtual perspectives. Due to its low cost of capture and display, it has been commonly used as immersive content for Virtual Reality (VR), and Mixed Reality (MR) applications [33], such as immersive movies. Although it could provide omnidirectional viewing experiences on 2D screens or Head-mounted Displays (HMDs), users are restricted to a limited Field-of-View (FoV) at each moment. Users may miss important events if they look in the wrong direction when watching a 360° movie. Since it is an unavoidable problem that users cannot perceive all information, attempts have been made to alleviate this issue by guiding the users to watch the eventful parts using visual indicators, redirecting view rotation, or displaying off-screen content. Visual indicators visualize the direction to regions-of-interest (ROIs) via symbolic diagrams [24, 4, 17, 16]. They effectively indicate where a target is, but lack visual content information. Navigation-based methods [24, 32, 27] change viewpoints automatically (auto-pilot) or inductively, forcing users to look in the direction of an important event. This method enables going through a series of events yet inevitably degrades immersion and prevents users from seeing multiple ROIs. The method relying on extra contents in [25], on the other hand, displays off-screen ROIs on the view window of a normal field-of-view (NFoV), which obscures some parts of the current scene.

As a pioneer work, Outside-In [25] proposed to use a set of 2D PIP windows to display off-screen ROIs. While having been demonstrated to outperform the conventional arrow-based navigation methods, Outside-In has the following drawbacks, limiting its ability to provide satisfactory visual experiences: (1) They use perspectively distorted windows for drawing PIPs to indicate the coarse positions of ROIs, which occupy a large area of the screen and may obscure important content of the main view window. Moreover, as the size of their PIPs remains constant, they often occlude and overlap each other. (2) The important content may not present significantly in PIPs, especially when the target object is too close to or too far from the PIP's camera, as its view range is constant. Some examples are shown in Fig.1.

This paper looks for a PIP method with more accurate recommended content and optimized presentation. Here, we focus on character-based 360° videos where ROI can be more clearly defined. For the other types of videos, such as scenery videos, users may be interested in exploring the whole scene, making it hard to define ROIs and the view direction guidance unnecessary. When playing character-

(a) Outside-In: Distant Character

(b) AdaPIP: Distant Character

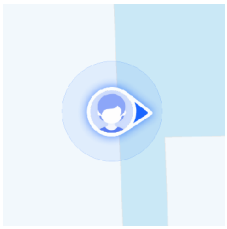
(c) Outside-In: Close Character

(d) AdaPIP: Close Character

Figure 1: Video play with Outside-In and AdaPIP. At the beginning of the video [14], a delivery man appeared from the corner. (a) Outside-In displays this event through a fix-range preview window, which is hard to be noticed. (b) AdaPIP adaptively reduces the context range of PIPs to visualize his movements clearly. (c) and (d): In the video [1], a man moves towards the door and is very close to the camera. The PIP of Outside-In can only show the upper body, while AdaPIP adaptively adjusts the view range to contain the whole object.

based videos, making the audience focus on ROIs containing characters' actions is essential for maintaining narrative drive. To better attract users to the ROIs, an appropriate PIP method needs to display useful contextual information for viewers to understand the content in the guidance window. Different levels of significance and view ranges of preview windows are required when guiding users to different characters. The direction indicator of a PIP window also considerably influences the effectiveness of view guidance. Conventional approaches such as the arrow-based guidance can be considered to be used in the PIP representation for ROI navigation, as they bring no additional learning cost, as shown in Fig.2.

Based on the above considerations, we present a new PIP-based guidance method providing a better 360° film-watching experience, namely AdaPIP. Our method mitigates the aforementioned issues of previous methods by introducing content-based adaptive PIPs with improved visual and interactive experiences. The basic element of AdaPIP is a circular plane focusing on the target characters with an attached arrow, which has been demonstrated to be an effective route-directing user interface (UI) in navigation applications [3]. Each off-screen character is previewed in a circular window, where an attached arrow indicates the direction and distance of the character. Furthermore, to alleviate the occlusion issue when there are multiple preview



(a) Route directing UI combination in AutoNavi

(b) Route directing UI combination in Google Map

Figure 2: Commonly used route maneuver user interface (UI) combinations: the circle represents the user's location, and directional elements like the arrow used in AutoNavi [3] or the isosceles trapezoid used in Google Map [11] point to the target direction.

windows, we adjust the size of PIPs according to the user's viewpoint and limit PIPs to an area in the lower middle of the main window where the important content is infrequent. In order to look for the optimal context range displayed in PIPs and accurately present the off-screen characters, we conduct user studies and demonstrate the following two claims:

- Users prefer more contextual information (larger view ranges) when the characters are smaller;
- Users prefer the characters to present in the same PIP window when they are close to each other.

Based on these observations, we develop an adaptive PIP method where the view range and included characters of a PIP window can be dynamically and smoothly adjusted with the guidance of content-related principles.

We further explored the applicability of PIP methods in a fully immersive environment by implementing AdaPIP and Outside-In in a VR headset. We conduct extensive experiments to demonstrate the effectiveness of our system in both 2D screens and VR environments by comparing AdaPIP with Outside-In and a baseline method where no directing technique is applied. Subjective ratings on several 360° video clips indicate the superiority of our method over Outside-In and the baseline method, demonstrating that AdaPIP provides more comfortable and effective watching experiences with a comparable degree of immersion. Furthermore, an extra test was conducted in both 2D and VR environments to show the benefits obtained by leveraging the adaptive mechanism for the content display of PIPs.

In summary, our contributions are as follows:

- A new Picture-In-Picture view guidance method, AdaPIP, with PIPs of content-aware adaptive sizes.

- An implementation of AdaPIP and related alternatives in HMD-based immersive VR environments.
- Comprehensive user experiments demonstrating the superior experience quality of AdaPIP on 2D screens and in VR environments.

The remainder of this paper is structured as: Sec.3 introduces the design of each element of AdaPIP. The adaptive scheme for dealing with different types of content is described and validated in Sec.4. Sec.5 details how we adapt AdaPIP and Outside-In to the VR environment. Sec.6 explain the experiments to evaluate AdaPIP. Sec.7 reports and analyzes the evaluation experiment results. Sec.7.2 summarizes the feedback from the participants.

2. Related Works

2.1. Attention Guidance in Virtual Environment

Considerable research has been done to explore the attention guidance techniques in AR and VR environments, as well as the particular case of virtual environment: 360° video. Rothe et al [35] divided these visual guidance techniques into two categories: on-screen guidance and off-screen guidance. On-screen guidance focuses on guiding users' fixation. By applying special screen effects (e.g., saliency modulation/ blurring/ stylistic rendering/ gaze direction), they guide users to focus on a specific part of the screen. However, this kind of instruction can only be seen when they are inside the viewers' current field of view. Therefore, on-screen guidance has significant limitations when users can freely choose where to look. On the contrary, off-screen guidance is dedicated to presenting off-screen content within the viewer's view range. Thus, we mainly focus on the off-screen guidance technique in our work.

A popular off-screen guiding method uses graphics and symbolic figures to indicate out-of-view targets. For example, arrows [24], haloes [4], radar points[17] and wedges [16] are adopted to provide spatial clues. For virtual environments that allow free movement, Adcock et al.[2] proposed composite wedge, 3D vector pairs, and a novel idea of rendering lit and shadowed areas to visualize the precise location of off-surface viewpoint in 3D space, thereby helping remote collaboration. On the other hand, instead of using graphics, a recent work, Outside-In [25] creatively introduced a picture-in-picture guidance method. It directly presents ROIs on small inline windows overlapping the main screen. More detail about Outside-In will be discussed later in this section.

In addition, the force rotation method also shows its effectiveness in ensuring that users catch all important events. This method rotates the scene until the ROI is inside the viewer's FOV. For example, Autopilot[24] automatically plans routes and directly brings the viewer to the position of the target when it is about to appear. Another example draws on the experience of traditional filmmaking. In traditional filmmaking, cutting can be used to show important details to the viewer. Pavel et al. [32] extended this experience in 360° video by delivering important areas to the viewer at every cut. However, it still needs to be investigated if the direction changes in the exact location due to the cuts will cause disorientation [35]. In a recent work [27], Liu et al. proposed a view-related playback method. They define several gaze conditions (e.g., looking at a specific ROI) and seamlessly loop the gate clips until the conditions are met. In this way, viewers have to turn to the given positions to see the important event. However, the looped audio will bring significant artifacts which greatly degrade user experience.

2.2. Outside-In

Outside-In is a visualization technique that uses spatial picture-in-picture previews to present the content of ROIs. Specifically, picture-in-picture is a widely used display method that introduces outside contents on the main screen via small inline windows. This method allows users to see the content out-of-view and allow them to decide whether to look at it. One disadvantage of this method is that the inline windows always overlap on the main screen, so important content can be blocked out. Another disadvantage is the missing information about the position of the ROIs [35], which has been solved in Outside-In by using the inline window itself as an arrow [25]. Inspired by the concept of perspective projection, Lin et al. placed the inline windows on the side near the ROI. They reshaped them according to their relative position, making them appear to have the right perspective relationship. In this way, users can naturally infer the positions of ROIs according to the appearance of PIP planes.

However, the PIP windows of Outside-In inevitably obscure the objects in the main window. To mitigate the occlusion problem, Lin et al. try to strike a balance by adjusting the PIP plane size according to the importance of the content behind it. Moreover, the inline windows only show a fixed view range, which can not fit different situations very well. For example, as shown in Fig.1, at the beginning of the video [14], a delivery man appeared from the corner, which is hard to be noticed in the PIP; in the video [1], a man moves towards the door and is very close to the camera, making the PIP only show his upper body. Plus, when two off-screen targets are close to each other, the PIP representing the farther target will cover a large part of the closer target's PIP. We address the above issue by adaptively presenting content and using a different layout.

2.3. Watching Experience of Head-Mounted Displays

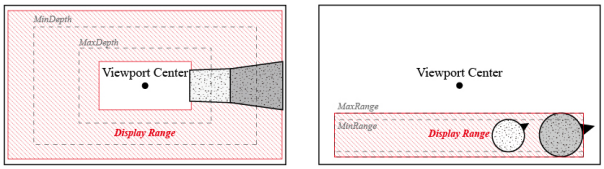
In comparison with a 2D screen, a VR headset like HMD provides a more immersive experience while watching 360° videos [5]. The heightened sense of immersion not only enriches the user's perception of presence, but also elicits a stronger emotional response to visually appealing content [10]. Yet, this advantage comes at the cost of increased symptoms of nausea, oculomotor and disorientation as illustrated in previous studies [17, 5].

VR headsets provide an authentic experience via the increment in both horizontal and vertical FOVs (≈ 80°-174° for horizontal FOV and ≈ 84°-114° for vertical FOV [18]) compared to a 2D screen. Drawing from the domain of visual perception, human vision may be divided into three principal regions: fovea, parafovea, and periphery. The fovea constitutes the central 2° of vision, whilst the parafovea encompasses a circumference of approximately 5° from the point of fixation. These two regions, collectively, are commonly referred to as central vision. Beyond the parafoveal area lies the peripheral region, commonly referred to as peripheral vision [22]. Central vision is responsible for perceiving high vision, shapes, and colors. However, peripheral vision is not accurate enough to perceive highly diverse visual content and is used for targeting the next eye movement [29]. The wider FOV of a VR headset also gives users more peripheral vision than a 2D tablet screen (e.g., VR headsets provide ≈ 90 degrees peripheral vision) [21, 23, 8] while 2D screens ≈ 30 × 20 degrees [31, 6, 9]). This creates sufficient room for displaying guide elements. By placing guide elements in peripheral areas, occlusion and interference issues can be mitigated while maintaining the ability to guide directions [28, 20].

3. Design of AdaPIP

The basic layout of AdaPIPs is shown in Fig.3. We display the off-screen targets on circular PIP preview windows inside the user's current view window. These windows are limited to the lower middle area, which can slide horizontally when the user continuously rotates their views or the off-screen target moves. Besides, an arrow is attached with a PIP to indicate the direction of the target object's position intuitively. We developed a panoramic video player with AdaPIP using the widely used game engine, Unity (version 2019.4.22f1c1). Our AdaPIP player can work on both HMDs (where users can turn their bodies or heads to explore the video) and 2D tablet devices (where users can click and drag the mouse to rotate their views).

The input to our video player contains the original 360° video along with several annotation files, including (1) manually specified characters and (2) the characters' tracking data (both spatial and temporal). The existing video tracking algorithm is not robust enough to provide sufficiently accurate object masks for 360° videos, especially when the videos contain cartoon characters or have poor lighting conditions. Since our work aims not to solve the tracking problem, we adopt a semi-manual annotation method to trace the path of the key characters, where we manually label the characters' positions at some keyframes and obtain their motion path via piecewise linear interpolation.



(a) Display range of Outside-In (b) Display range of our method

Figure 3: A comparison of the display range between Outside-In and Our Method.

In order to lessen occlusion caused by PIPs, we choose to render preview windows on the lower part of the view window. When watching the video using a 2D display, users' sight is usually perpendicular to the screen. Thus we superimpose the PIP image planes on the 360° video. In a VR environment, PIP planes are designed to rotate about the user to remain perpendicular to the user's view direction. In addition, we set the distance between the view plane of PIPs and the user to 0.3 meters to support possible real-time interactions since this distance can be easily reached in a VR environment.

In the following subsections, we first introduce how we determine the sizes and positions of a PIP preview window and its arrow to indicate the distance and direction of an object intuitively. Then we describe how our preview windows react to relative position changes between the user's viewpoint and the target objects.

3.1. Distance Representation

Using the position of the PIP window is an intuitive way to indicate how far the target object is from the user's current viewpoint. We thus make PIP windows slide horizontally in the specified region when the user or the off-screen target moves. Given a 360° video represented by equirectangular projection, we use latitude and longitude to define the unique position on a frame, where latitude ranges from -90° to +90°, and longitude ranges from -180° to +180°. As shown in Fig.4, assuming that the current viewport center is V, the position of the PIP on the screen is P. The character outside the current FOV is C, and the distance between P and C can be defined via normalized Euclidean distance, which is a value between 0 and 1:

$$D = \sqrt{((\frac{\Delta latitude}{90°})^2 + (\frac{\Delta longitude}{180°})^2)/2} \quad (1)$$

To avoid occlusions when multiple PIP windows have similar relative distances between the user and the contained target object, we constrain the distance between two PIPs to be greater than a threshold d_min.

Apart from the position of a PIP, its attached arrow can also indicate the relative distance as a complementary. When a user rotates their head, and the screen center moves away from the off-screen target, the distance between the arrow and the PIP center is set to increase accordingly, which looks like being stretched. The arrow is gradually pulled back to the PIP window when the user's view center approaches the target. Specifically, the length of an arrow L is linearly decided by the distance D and the predefined maximum/minimum arrow lengths $L_{max/min}$:

$$L = L_{min} + D \times \frac{L_{max} - L_{min}}{D_{max}} \tag{2}$$

Please see our supplement video for how the PIPs work when users watch 360° videos.

3.2. Direction Representation

To represent the rotation direction of the target object, we rotate the attached arrow about the PIP center by the angle between the view direction and the current direction of the target object. As shown in Fig.4, P denotes the position of the PIP on the user's view window, and C denotes the position of the off-screen target. The direction of the vector $\overrightarrow{PC}$ is used as the direction of the attached arrow. This kind of indication method is often seen in navigation applications [3, 11] and has been shown to be effective in reducing the learning cost for users. Also, since we use the arrow instead of using the PIP itself to indicate the location of the target as in Outside-In [25], our PIP window no longer needs a large display area, which can mitigate the occlusion issue. See Fig.3.

3.3. View-Dependent Interactions

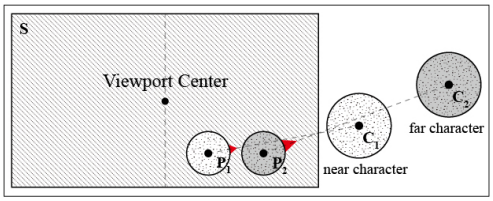
To provide a better watching experience, our PIP windows can make real-time interactions according to the user's current view direction and FOV.

Display Visibility Our interaction scheme works when the user starts watching the 360° film. If the user is not looking at a specific character, the PIP for that character pops out. After the user turns their head to that character, the PIP fades out. In addition, we also implemented autopilot interactions. By clicking any PIP, the user's view can directly turn to the corresponding direction for the target character.

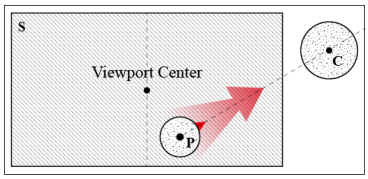
Adaptive Scaling When the user changes their viewing direction, we dynamically adjust the size of all active PIPs using the angle between the current and the direction of the target characters in real time. We use the distance to the nearest target to determine the window size S of all the PIPs by:

$$S = \begin{cases} S_{min}, & D < D_{lower} \\ S_{max}, & D > D_{upper} \\ S_{min} + (D - D_{lower}) \times \frac{S_{max} - S_{min}}{D_{upper} - D_{lower}}, & \text{otherwise} \end{cases} \tag{3}$$

where S_{min} and S_{max} denote the minimum and maximum size of the PIPs, which are set to 30 and 64 respectively. D_{lower} and D_{upper} are the two thresholds for the distance to the characters to determine whether the minimum or maximum window size should be applied.



(a) Distance Representation



(b) Direction representation

Figure 4: The direction and distance representation in AdaPIP: (a) The distances between the enclosed characters and the user are represented by the distance the PIP deviates from the viewport center. In other words, when the user approaches the character outside the screen, the PIP will be closer to the viewport center. (b) We define the direction of the attached arrow as the direction from the center of the PIP to the center of the character.

4. Adaptive Context

In previous works such as Outside-In, an off-screen ROI on a PIP plane is rendered with the same FOV as the main window [25]. It causes serious issues when characters inside an off-screen ROI are too far or too close from the viewpoint. The characters may be too small to be observed when it is far away from the users' viewpoint and too large to be displayed entirely when close to the viewpoint. Since the PIP windows should focus on the characters rather than the entire ROI areas, we can prompt the off-screen characters more efficiently by adaptively adjusting the view range of the content rendered in PIPs. However, no previous research has been done to reveal users' preference for the view range of PIPs; we designed the following two research

questions and performed two experiments to gauge whether users have clear preferences.

RQ1: When watching 360° videos with PIP prompts, do users have a preference for the view range of content in PIPs? More specifically, do users prefer a wider range with more contexts or a narrower range?

RQ2: When there are multiple characters close to each other, do users like them to be shown in the same PIP or separately?

4.1. Study for View Ranges

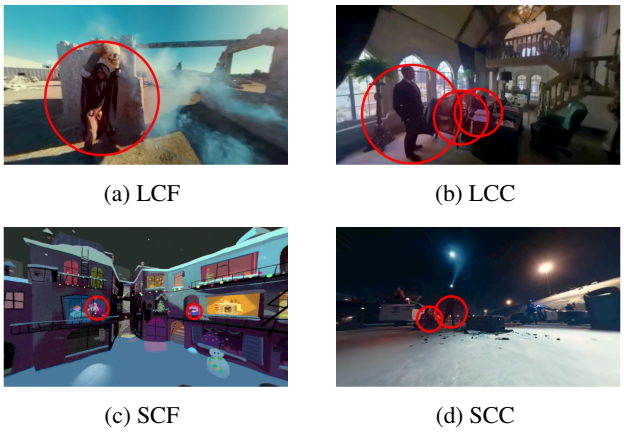
4.1.1 Experiment design

We collected 6 character-based 360° videos from YouTube [13, 7, 19, 34, 14, 12]. Then we extract 8 video clips of 10-15s from the above 6 videos according to the size of the characters and the relative distance between the characters. These 8 video clips can be divided into two groups of 4 clips according to the size of the characters: large character video clips (LC) and small character video clips (SC). In the LC video group, there are 2 video clips (LCF) with two characters far apart, and 2 video clips (LCC) with two characters close together, the same for the SC video group (SCF and SCC). See Fig.5. We use video clips with a length of 10-15s because character sizes and relative distances between characters vary rapidly across all videos, making it difficult to find long clips where character sizes and relative distances remain stable. We use a circular bounding box to represent the character and take the mass center of the circle as the character center, see Fig.6. The character's moving path is recorded as the path of its bounding box via a semi-manual annotation process which we introduced in Sec.2.

We designed two experiments to answer the two research mentioned above questions. In the experiment for RQ1, we let each participant watch two LCF videos and two SCF videos. We randomly assign narrow PIPs and wide PIPs to the played videos, where a narrow PIP displays a 10% larger area than the character's bounding box, and a wide PIP displays a 60% larger area. In the experiment for RQ2, we provide 2 LCC and 2 SCC video clips to the participants. We asked each participant to watch each video clip twice, where PIPs show grouped characters or each of the characters separately. For PIPs displaying grouped characters, we merge the characters that have intersections between their bounding boxes and display them on a single PIP window; For separate PIPs, we assign a PIP window for each character.

4.1.2 Procedure and Measures

We recruited 10 participants (6 male, 4 female) for the two experiments. The participants were all college students aged 19-28, and 7 had previously watched 360° videos via



(a) LCF (b) LCC

(c) SCF (d) SCC

Figure 5: Red circles show the characters in the videos. According to the size of the characters and the relative distance between the characters, video clips are divided into the following four types: (a) LCF: videos with characters that are large and far away from each other. In this frame, a character shoots at another character in the opposite direction and out of view. (b) LCC: videos with characters that are large and close to each other. (c) SCF: videos with characters that are small and far away from each other. (d) SCC: videos with characters that are small and close to each other. Videos are from [13, 7, 19, 12].

2D screens. We designed two experiments for the two research questions accordingly. At the beginning of each experiment, a brief tutorial was provided to participants, explaining the PIP-based guidance method used in this experiment (wide range and narrow range PIPs for the first experiment, PIPs with grouped characters and separate characters for the second experiment). Participants were encouraged to try different view directions to understand how AdaPIPs work when watching the test videos. After the tutorial, participants were asked to watch 360° videos with different PIP-based guidance methods. They were told which video type (LC or SC) and which method would be shown before watching. For each experiment, participants were required to watch 4 video clips, so each participant needed to watch $2 \times 4 = 8$ video clips throughout the study. The order of the videos to play is randomized for different participants. After watching a group of videos, participants were asked to rate the watching experience using a score between 1 (worst) and 7 (best).

This experiment was conducted using a 17" 1920×1080 laptop screen. The size of the play window is 1778×1000 pixels. Participants can click and drag the mouse to adjust the viewing direction and click the PIP to turn to the corresponding off-screen character.

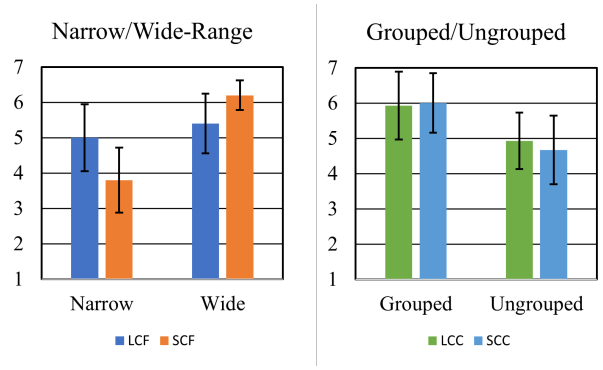


Table 1: Average ratings for narrow/wide-range (left) and grouped/ungrouped methods (right). Error bars show standard deviations.

4.1.3 Results

Context Range In experiment 1, we collected 10 (participants) $\times$ 2 (methods) $\times$ 2 (video types) = 40 ratings. For videos with large characters, the rating of narrow range PIPs (μ = 5, σ= 0.943) and large range PIPs (μ = 5.4, σ= 0.843) don't get much difference. For small-character videos, we find that participants prefer a large range (μ = 6.2, σ= 0.422) to narrow range (μ = 3.8, σ= 0.919). See Tab.1.

We further performed a two-way repeated-measures ANOVA. There was a statistically significant interaction between ranges and video types, where $F(1, 9) = 45$ and $p < 0.05$. We analyzed the effect of ranges at each video with adjusted p-values using the Bonferroni multiple testing correction method. A significant effect of different ranges was found for videos with small characters ($p = 0.0000512$) but not for large character videos ($p = 0.686$). The pairwise comparisons also showed a significant difference between the ranges for small character videos. It demonstrates that users have no obvious preference for the display range for large characters, and users prefer a wide context range for small characters.

Grouped or Separate Characters Same as in experiment 1, we got another 40 sets of ratings in experiment 2. We found that all participants gave grouped characters higher scores. Some expressed that grouping characters creates fewer distractions and helps them see the interaction among characters, attracting them to watch the corresponding event using their main view window. As indicated in Tab.1, for large-character videos, using PIPs with grouped characters (μ = 6.1, σ= 0.568) got a higher score than ungrouped (μ = 3.7, σ= 0.675), same for the small-character videos (grouped: μ = 6.2, σ= 0.422, ungrouped: μ = 2.6, σ= 0.843). We further analyzed the results with a two-way repeated-measures ANOVA, where a statistically significant interaction between the methods and the videos is found with $F(1, 9) = 36$ and $p < 0.05$. Therefore, the effect of both the grouped and ungrouped methods was analyzed in each video, and a significant effect was found for both large-character videos and small-character videos. We also analyzed the effect of videos on each method and only found a significant effect for the ungrouped method. The pairwise comparisons further illustrated a significant difference between LC and SC videos for the ungrouped method. In conclusion, for any type of video, statistically, significant differences indicate that prefer the characters to be grouped when they are close. Moreover, users rate the ungrouped method worse for SC videos.

(a) Bounding Box (b) Narrow Range (c) Wide Range

Figure 6: (a) shows the bounding box for the character. (b) For the narrow range, PIP displays a 10% larger scope than the character's bounding box, and (c) a wide range PIP displays a 60% larger scope. This video frame is from [15].

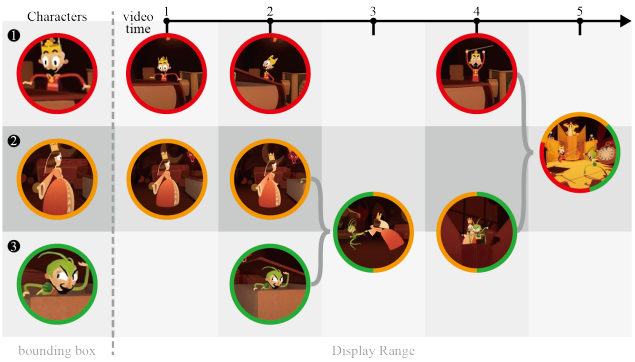


Figure 7: Illustration of the adaptive context mechanism. Smaller characters have more contextual information rendered on the PIPs; larger characters have less contextual information; when the characters' bounding boxes intersect, we consider these characters to have possible interactions in that frame and display them on a single PIP window. For example, at time point 2, character 2 and character 3 are running together and have intersections in their bonding boxes, thus they are displayed in one PIP window. Video Frames are from [15].

4.2. Content-Aware Context Range

The above experiments found that the user's preference for context range has a strong link with the sizes and positions of characters. Therefore, we apply a three-stage process to dynamically adjust the position and the context

range of PIPs based on the experimental results.

First, We count the number of interactions among characters and group the characters in different time ranges since users clearly prefer whether the characters should be grouped when they have close relationships. When the characters' bounding boxes intersect, we consider these characters to have possible interactions in that frames. If the duration of the intersection exceeds a specified threshold (η=10), these characters are considered to have a real relationship. When calculating the duration, we allow the characters to be separated for a short period (shorter than η), as long as these characters still exist in the picture, ensuring the relationship's temporal stability. Suppose some characters have a relationship in a certain period, we consider them to belong to the same context group, and we only use one PIP to display these characters. If a character does not interact with other characters, it forms its group.

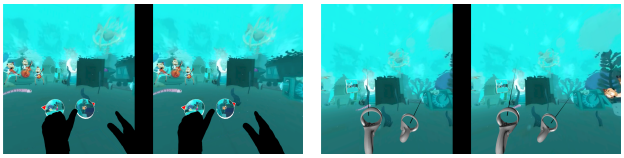
Second, we calculate the center position and context range of PIPs based on the bounding boxes of the character groups in each frame. We use the narrow range for a group with a single small-size character. If multiple characters are in one group, we perform a weighted average to calculate its center.

At last, we check how the groups change over time and make a smooth transition between different context ranges. For example, before and after the merging or separation of groups, there will be noticeable context range and character size changes. In order to ensure the smoothness of those changes, the context range of the transition frames will be interpolated by Laplace Smoothing by the original ranges before and after the transition.

Applying AdaPIPs reduces the number of needed PIPs since we consider the characters group-wisely. Also, compared with only using a narrow context range, we provide the necessary background and interaction information. We also enable an AdaPIP window to adapt to the size change of characters dynamically. When a character moves towards the camera, the view range will increase so that users can notice the distance change of the character.

5. Adaption to VR Environment

For a more comprehensive assessment of the AdaPIP method, we also explored how to adapt PIP technologies in a VR environment in addition to 2D-screen-based 360° video play. Video viewing with HMDs provides a wider FOV, as well as a wider peripheral vision [18, 21, 36, 8]. The peripheral area can be used to effectively display guide elements while mitigating occlusion and interference issues [20, 28], which indicates an enlarged displaying area for the prompt windows when adapting Outside-In and AdaPIP into VR environments. Moreover, both Outside-In and AdaPIP superimpose prompt windows on the original 360° video when played on 2D screens. We put each PIP on a plane at different depths from the 360° video to inherit that idea. We also enabled stereoscopic display for a higher sense of perceived depth [37].



(a) AdaPIP (b) Outside-In

Figure 8: The binocular view of the user interface of AdaPIP and Outside-In in VR environment: (a) The VR version of AdaPIP presents controllers as a pair of hands. When a user touches the circular PIP plane with their "hand," they jump to the corresponding perspective. (b) In the VR version of Outside-In, the controller emits a black ray. When the ray is aimed at the picture-in-picture plane, the user can choose to press the trigger button, at which point the black ray will turn green and trigger autopilot. This video is from [15]

AdaPIP In a 2D scenario, users can make an autopilot to the view direction for an ROI by clicking the corresponding PIP. We also enable users to trigger an autopilot by using the controller to "click" the PIP window in VR, see Fig.8. Besides, a pitch rotation of the view direction in a VR environment may confuse users' navigation. For example, when a user makes an autopilot to the sky, the physical head direction may remain straight ahead. If the user looks down, they will see the object in front of them instead of the ground. Therefore, we limit the pitch rotation and only allow yaw rotation when autopilot happens in VR. We indicate the pitch angle to the ROI center by a stretched arrow after the autopilot.

Outside-In Outside-In places the PIP windows around the screen center in 2D [25]. In the VR environment, we adopted the same method and limited the display range of the PIP windows to the peripheral area. Unlike AdaPIP, the depth of each PIP plane of Outside-In implies the distance to the corresponding object, which linearly decreases with the distance between the target object and the view center. We use a large depth range to achieve a similar appearance as in the original 2D Outside-In. However, that makes the PIP plane too far from the user to be reached for autopilot. To solve this issue, a virtual ray emitted from the controller is used to hit the PIP plane, and the user can press the controller's button to trigger an autopilot, see Fig.8. To avoid the aforementioned navigation confusion, we also limit the pitch rotation and use the PIP's position to indicate the needed pitch rotation after the autopilot.

Video name	Genre	Author	Length(sec)	
			Clip1	Clip2
Back To The Moon [15]	Comedy	Google Spotlight	51	66
Help [13]	Horror	Google Spotlight	53	49
Knives [1]	Thriller	Indie	57	51
Lions [30]	Documentary	National Geographic	66	70

Table 2: Summary of example videos.

(a) 2D screen test (b) VR test

Figure 9: Setup for the (a) 2D screen test and (b) VR test.

6. User Experiment Settings

To test whether our approach improves user experiences, we conduct a user study comparing AdaPIP with Outside-In and a baseline method where no PIP guidance is provided in 2D and VR environments. We collected another set of 4 videos covering a variety of genres from Youtube to demonstrate the generalizability of our method for different narrative types, with the detailed information provided in Tab.2. Since our method focuses on the characters in videos, we did not use scenery videos. We divided each video into two discontinuous video clips with a duration ranging from 49 seconds to 70 seconds and randomly labeled each video clip as 1 or 2 as clip 1 always displays in a 2D environment and clip 2 in a VR environment.

6.1. Method

To test the effectiveness of our method in 2D and VR environments, our user study includes two formal tests: **2D screen test** and **VR test**. Participants were asked just to take one of these two tests. Before the formal test, participants were given a brief introduction to all methods and interaction schemes included in our experiments. They were allowed to experience these techniques in both 2D and VR while watching test videos that were not included in later tests until they got familiar with different methods. After the formal test, an **extra test** was conducted to explore further whether AdaPIP can help participants recognize the prompt content.

2D screen test For this test, participants were asked to watch 4 video clips 3 times via a desktop monitor and informed of the PIP technique. One of the following 3 methods was applied for each time: baseline, Outside-In, and AdaPIP. The baseline was always applied first to compare Outside-In and AdaPIP directly. Outside-In and AdaPIP were randomized and counter-balanced, as was the order of the video clips presented to the participants. Participants were asked to fill out a questionnaire after watching one video with different PIP methods. They were also asked to rate 3 methods in terms of Q1: overall performance and Q2: understanding the level of spatial relationship. Besides, they were asked to rate Outside-In and AdaPIP in terms of Q3: interference level and Q4: recognition level of the prompt content.

VR test Participants need to wear an HMD with two controllers for the VR test. They were asked to watch another 4 video clips for 3 times with different PIP techniques. The orders of the 3 methods and the video clips are the same as the 2D test. After watching a video 3 times, participants removed the HMD and took a break while filling out a questionnaire and rating their experience using the same criteria as the 2D test.

Extra test We conducted an extra test to validate whether the adaptive mechanism can help participants recognize the prompt content. We compared AdaPIP with its non-adaptive version, which uses the same interface as AdaPIP but displays a fixed range of content on the PIP plane.

After the formal test, the extra test was presented to each participant. Participants who took the 2D screen test during the formal testing session were asked to watch the same 4 videos they saw in the 2D screen test through a desktop monitor; similarly, participants who took the VR test were asked to watch the same 4 videos as the VR test wearing an HMD. All participants were asked to watch each video twice, applying AdaPIP or its non-adaptive version in random order. The order of video clips is also randomized and counter-balanced. After watching a video twice, participants need to rate the recognition level of the prompt content for the two PIP methods.

6.2. Participants

We recruited 28 (16 males, 12 females) university students from different majors as our participants, aged 18 to 26. 13(8 males, 5 females) of them signed up for the 2D screen test, and the remaining 15(8 males, 7 females) signed up for the VR test. For participants taking the 2D screen test, 7 of them had watched 360° video via 2D screen before; for participants taking the VR test, none of them had worn an HMD to view 360° video in the past.

6.3. Apparatus

For the 2D screen test, we used a 17-inch laptop HD monitor with an Intel Core i7 processor and an NVIDIA GeForce GTX2070s graphics card. The distance between the participants and the monitor is about 40cm. We built our platform in Unity (Version 2019.4.22F1C1) and played 360° videos via Unity's Play Window (1778 ×1000 Pixel). Participants can click and drag the mouse to rotate the view or click the PIP window to jump to the corresponding view.

We used an Oculus Quest 2 connected to the laptop used in the 2D screen test for the VR test. The Oculus Quest 2 has a single eye resolution of 1832×1920 pixels with a horizontal FOV of 89 degrees (+-4) and a vertical FOV of 93 degrees (+- 5.1°) [18]. We play 360 videos on the Unity platform and stream them to Oculus Quest 2. A swivel chair was provided to participants, and they were able to rotate in the yaw dimension easily. Participants were asked to use two controllers while watching the video. They can jump to the corresponding view by touching the PIP window via controllers.

6.4. Measurements

After viewing each video clip, participants were asked to rate the guidance method using a 7-point Likert scale (1 lowest, 7 highest). At the end of each formal test, we conducted a brief interview about how they assessed the assistance of each technique and what aspects they liked or disliked. The full 2D screen test, including practice, interviews, and the extra test, lasted about 40-50 minutes; the whole VR test lasted about 50-60 minutes.

7. Results

7.1. Subjective Rating

7.1.1 2D Screen Test

For the 2D screen test, we collected the ratings from 13 participants in terms of the aforementioned criteria.

Q1: Overall Performance According to the results presented in Tab.3, we can see that For all test videos, participants ranked AdaPIP as the preferable method (1-least preferable, 7-most preferable) in terms of overall performance. We further performed a two-way repeated-measures ANOVA and found a statistically significant interaction between different methods, $F(2, 24) = 17.334$, $p < 0.0001$. No statistically significant interaction was found between different videos and between methods and videos. Pairwise t-test comparisons demonstrated significant differences between methods.

Q2: Understanding Level of Spatial Relationship As shown in Tab.3, most participants gave higher scores for AdaPIP, believing that AdaPIP can help them efficiently find the position of characters in 360° space and understand spatial relationships. A two-way repeated-measures ANOVA and pairwise paired t-test comparisons were performed, and a statistically significant interaction between different methods was found with $F(2, 24) = 54.194$ and $p < 0.0001$. No statistically significant interaction was found between different videos, or between methods and videos.

Q3: Interference Level For the ratings of interference level, higher scores represent higher levels of interference (1-least interference, 7-most interference). As shown in Tab.3, the interference level of AdaPIP is significantly lower than Outside-In for all videos. Two-way repeated-measures ANOVA and pairwise paired t-test comparisons were also performed, and a statistically significant interaction between different methods was found with $F(1, 12) = 39.103$ and $p < 0.01$. No statistically significant interaction was found between different videos or between methods and videos.

Q4: Recognition Level of the Prompt Content AdaPIP and Outside-In showed a similar level of readability of the content prompted by the PIPs; see Tab.3. We did not find significant interactions between the methods, between the videos, or between the methods and the videos.

7.1.2 VR Test

For the VR test, we collected the ratings from 15 participants in terms of the criteria mentioned above.

Q1: Overall Performance In the VR environment, most participants had the highest preference for AdaPIP and the lowest preference for baseline in terms of overall performance. See Tab.4. We also performed a two-way repeated-measures ANOVA and pairwise paired t-test comparisons. There was a statistically significant interaction between different methods, $F(2, 28) = 50.984$, $p < 0.0001$. No statistically significant interaction was found between different videos, and no statistically significant interaction was found between the methods and videos.

Q2: Understanding Level of Spatial Relationship As shown in Tab.4, most participants felt that AdaPIP could provide effective instructions and thus help them understand the spatial relationships among the characters in the videos. Two-way repeated-measures ANOVA and Pairwise paired t-test comparisons were performed. A significant interaction between different methods was found with $F(2, 28) = 24.702$ and $p < 0.0001$. No statistically significant interaction was found between different videos, and no statistically significant interaction was found between methods and videos.

Q3: Interference Level Similar to the 2D screen test, for evaluating the interference level, higher scores represent higher interference. As shown in Tab.4, most participants rated Outside-In higher, feeling that Outside-In caused more distractions when watching videos. No statistically signifi-

	Back to the moon						Help						Knives						Lions					
	Baseline		Outside-In		AdaPIP		Baseline		Outside-In		AdaPIP		Baseline		Outside-In		AdaPIP		Baseline		Outside-In		AdaPIP	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Q1	4.69	1.030	5.38	0.768	5.92	0.641	3.62	1.260	4.92	1.040	5.54	1.050	4.23	1.240	5.15	0.899	6.08	0.862	3.85	1.520	5.15	1.280	6.08	0.494
Q2	3.62	0.961	4.69	0.947	5.62	1.120	3.38	0.650	4.85	1.070	6.00	1.080	4.23	1.010	5.38	0.870	6.23	0.725	3.69	1.030	4.62	1.450	5.54	1.390
Q3			3.31	1.600	2.46	1.560			3.92	1.550	2.31	1.320			3.31	1.650	2.31	1.700			3.62	1.610	2.23	1.240
Q4			5.08	0.862	5.54	1.050			6.08	0.641	5.54	1.270			6.00	1.080	6.00	1.000			5.46	0.877	5.46	0.776

Table 3: Mean and standard deviations of the ratings in the 2D screen test. For Q1, Q2, and Q4, 1 means the least preferable and 7 means the most preferable; for Q3, 1 is the most preferable, since a less inference level means a better experience.

	Back to the moon						Help						Knives						Lions					
	Baseline		Outside-In		AdaPIP		Baseline		Outside-In		AdaPIP		Baseline		Outside-In		AdaPIP		Baseline		Outside-In		AdaPIP	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Q1	3.93	1.220	4.80	0.941	5.87	0.834	3.47	1.460	5.13	1.060	6.07	0.884	3.00	1.560	4.87	0.915	6.00	0.655	3.80	1.260	4.87	1.120	5.87	0.743
Q2	3.87	1.460	5.53	1.190	6.33	0.900	3.67	1.990	5.20	1.210	6.00	0.756	3.73	2.340	5.33	1.290	6.07	0.704	3.73	1.940	5.07	1.030	5.73	0.704
Q3			3.80	1.320	2.27	0.884			3.53	1.190	2.20	0.862			3.07	1.100	1.93	0.799			4.07	1.390	2.67	0.900
Q4			4.47	1.640	5.53	1.460			5.00	1.690	5.40	1.400			5.33	1.230	5.53	1.190			5.33	1.110	5.47	0.915

Table 4: Mean and standard deviations of the ratings in the VR test. For Q1, Q2, and Q4, 1 means the least preferable and 7 means the most preferable; for Q3, 1 is the most preferable, since a less inference level means a better experience.

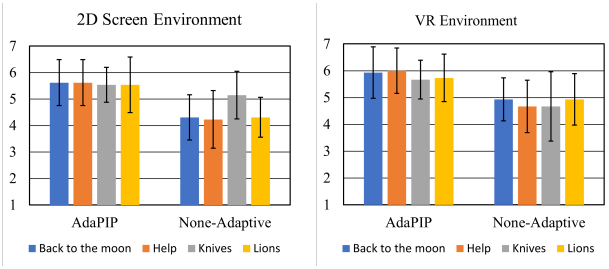


Figure 10: Average ratings for the extra test in 2D screen environment (left) and VR environment (right). Error bars show standard deviations.

	Back to the moon				Help			
	AdPIP		None-Adaptive		AdPIP		None-Adaptive	
	mean	sd	mean	sd	mean	sd	mean	sd
2D Screen	5.62	0.870	4.31	0.855	5.62	0.870	4.23	1.090
VR	5.93	0.961	4.93	0.799	6.00	0.845	4.67	0.976
	Knives				Lions			
	AdPIP		None-Adaptive		AdPIP		None-Adaptive	
	mean	sd	mean	sd	mean	sd	mean	sd
2D Screen	5.54	0.660	5.15	0.899	5.54	1.050	4.31	0.751
VR	5.67	0.724	4.67	1.290	5.73	0.884	4.93	0.961

Table 5: Means and standard deviations of the ratings for different video clips in the extra test.

cant interaction was found between different videos or between methods and videos. We also performed paired t-test comparisons, showing that the scores of different methods were significantly different.

Q4: Recognition Level of the Prompt Content As revealed in Tab.4, most participants thought that there was no significant difference between AdaPIP and Outside-In in terms of the recognition level of the prompt content. They said both of them could effectively help them understand the plot. A two-way repeated-measures ANOVA was performed, and no significant interaction between different methods, between the different videos, or between the method and the video.

7.1.3 Extra Test

2D Screen Environment We collected ratings from 13 participants for the extra test in a 2D environment. According to the results presented in Tab.5, most participants indicated that adaptive content can more clearly present the actions of characters, and removing the adaptive scheme lessens the recognizability of the prompt content. Therefore, they gave the former higher scores, see Fig.10.

A two-way repeated-measures ANOVA was performed, and a statistically significant interaction was found between the methods and videos with $F(3, 36) = 3.220$ and $p < 0.05$. Therefore, the effect of the method variable was analyzed in each video. P-values were adjusted using the Bonferroni multiple-testing correction method. The effect of treatment was significant for the video Back to the moon, Help, Lion, but not for the video Knives. Pairwise com-

parisons, using paired t-tests, show that the mean score was significantly different between AdaPIP and NoadaPIP for the video Back to the moon, Help, and Lion but not for the video Knives. This suggests that there is no significant difference between AdaPIP and its non-adaptive version for video Knives. However, for the other three videos, there is a statistically significant difference in the scores between the two methods. By inspecting the video "Knives"[1], we found that its character sizes are moderate and keep nearly constant. Therefore, AdaPIP's results are similar to non-adaptive PIP methods only in this video.

VR Environment We collected ratings from 15 participants for the extra test in a VR environment. Most participants also felt that the adaptive scheme of AdaPIP could improve the recognizability of the content presented on the PIP planes. As indicated in Tab.5 and Tab.10, AdaPIP outperforms the non-adaptive method. We further performed A two-way repeated measures ANOVA and found a statistically significant interaction between different methods with $F(1, 14) = 75.162$ and $p < 0.0001$. There is no statistically significant interaction between different videos, or between methods and videos.

7.2. Interviews

We recorded the interview conversations in the form of audio recordings and excerpted several answers in this section. Among the 28 participants we recruited, Participant 1 to Participant 13 took the 2D screen test (hereafter referred to as P1 to P13), and P14 to P28 participated in the VR test.

7.2.1 Overall Preference

In a 2D environment, 10 people felt that adding PIP windows to 360 videos can enhance their watching experience, while 3 other people prefer watching without guidance. As P3 claimed, "I think watching without guidance is a natural way of viewing videos, with no additional cognitive load." 9 out of 13 said they prefer AdaPIP more than Outside-In. P8 said, "AdaPIP uses a familiar UI that I have experienced in video games." The remaining 4 participants expressed their preference for Outside-In. "Outside-In feels like surveillance windows." P6 says, "With these windows, I can monitor every event in all directions."

While watching with VR headsets, 13 participants claimed their preference of watching with guidance. 2 participants thought the necessity of adding PIP windows depends on the content of the video. For videos like Back to The Moon[15], changes in light and scenery can effectively guide participants to focus on the protagonist, and it is no need to add extra guidance. However, for videos like Help[13], protagonists are constantly moving from side to side at high speed throughout the video, in this case, PIP windows are needed. In addition, 8 participants said that despite having a swivel chair, they still disliked turning their heads or bodies while watching the video. It really enhances their viewing experience if they can see any plot without turning their heads or bodies. 13 participants expressed their preference for AdaPIP; 1 participant said he had no particular preference; P18 said he prefers Outside-In because "Outside-In displays out-of-view content in a larger window and is easier to recognize."

7.2.2 Spatial Guidance

26 out of 28 participants said AdaPIP was more effective in guiding direction. 5 participants said that the attached arrows in AdaPIP provided easy and efficient directions and helped them find targets faster. Meanwhile, P18 added, "The red arrow on the PIP window makes me want to jump to the indicated viewpoint." P8 also had a similar opinion, "I think AdaPIP provides me with a powerful incentive to explore the prompt content". Nevertheless, P24 said, "In the VR environment, the PIP windows of AdaPIP are displayed really close to me. There were some cases where I ignored the PIP windows when I changed my fixation to the video behind me. For Outside-In, the PIP windows are usually far away from me, and I am less likely to ignore them. That's why I think outside-in performs better."

7.2.3 Context Range

Among the 28 participants, 8 participants felt that Outside-In's PIP window was larger, so it displayed clearer and more comprehensive content; 18 participants thought there was no significant difference between AdaPIP and Outside-In in terms of the legibility of the content displayed in the PIP window; 2 participants said AdaPIP's PIP window could display more clear content. Plus, there are 6 participants who expressed their preference for the adaptive range scheme in AdaPIP. P1 suggested, "While Outside-In's PIP windows can display clearer content, I sometimes see extra stuff in the PIP window. For example, in a situation where one character on the left is talking, an arm of another person appears on the other side." P16 said, "I love the idea of adaptive content because it delivers more precise information."

In addition, 5 participants mentioned the size of characters in PIP windows. P26 said, "It's strange that the character's size in AdaPIP seems to be different from its original size in the video. For example, a character in a PIP sometimes looks large, but he is actually smaller in the original video." "But it does not affect my understanding of the plot," she added.

7.2.4 Interference

26 of the 28 participants reported lower interference levels with AdaPIP compared to outside-in. Two participants thought there was no significant difference. P1 said: "Outside-In has larger PIP windows, which reduces immersion. Sometimes these windows can severely obscure the video behind, which annoys me." P10 thought, "When there are multiple targets, the PIP windows in Outside-In can easily overlap each other." P24 shared her thoughts on the watching experience in a VR environment: "Since the AdaPIP's prompt window is displayed below my sight and very close to me, it is less disturbing when I am focused on the video behind. However, Outside-In's PIP plane is close to the video, and it's distracting while watching the video."

7.2.5 Interactions In VR

10 of 15 participants preferred the interaction method of AdaPIP for triggering autopilot. P13 said, "AdaPIP's circular window can be touched with controllers, which is a novel experience that makes me feel immersive and has a stronger sense of interaction." 8 participants said that while this was a novel interaction, having to raise their hands every time for autopilot would make them feel tired. According to P18, "Outside-In has rays that the controller emits to the target. While these rays are somewhat distracting, this interaction is easier to perform." "I can do it by just putting my hands on my lap and pressing the trigger button." P24 said.

7.3. Discussions

From the above results, it can be seen that 1) In both 2D and VR environments, users always get a better viewing experience with the help of PIP guidance; 2) Our method is better evaluated compared to outside-in. On the one hand, our method can effectively guide users to find targets and improve their understanding of spatial relationships. On the other hand, compared to Outside-In, our method can effectively reduce the occlusion problem and has a lower interference level in both 2D and VR environments. Furthermore, the extra test shows that our method can prompt more accurate content by adopting the adaptive context range. The only exception in our experiments is the video Knives[1], where the characters maintain a modest and constant size. For most videos, our adaptive scheme can effectively improve the recognizability of the content in PIPs.

We provided two different interaction modes in the VR environment, virtual hands or emitted rays for AdaPIP and Outside-In, respectively. As can be seen from the results and interviews, participants felt that the two types of interactions had both advantages and disadvantages. Using virtual hands to touch the PIPs in AdaPIP has a lower level of interference and provides a novel experience, but it requires frequent hand movement. Outside-In's ray-based interaction mitigates the issue of fatigue but causes more interference problems.

In all, our method has a better overall performance than Outside-In, providing users with a better viewing experience in both 2D screens and VR environments.

8. Limitations and Future Works

8.1. Limitations

Our AdaPIP is designed to work with films of character-based stories. For videos that were captured for users to explore freely, such as scenery videos, our method is not applicable. Besides, there are distraction issues. Some users mention that they would like to change their viewpoint when PIPs pop up because they may consider the arrow of a PIP as a hint to change their viewpoint. Thus, the viewer may be less patient in watching the content of PIP preview windows. However, we also believe that in most 360° videos, making users desire to change their viewpoint can encourage users to explore the 360° virtual space fully.

8.2. Future Works

Automatic Labeling and Tracking In our work, characters are manually annotated for each video to ensure PIP previews' accuracy. The process requires labor-intensive work, especially for long videos. This step can be replaced by leveraging video segmentation, and object tracking methods [38, 26]. Since our algorithm takes almost negligible time, our method can be easily integrated with a video play application to provide a smooth PIP experience if the characters can be identified and tracked in real time.

Importance Suggestion The size of our PIP windows always keeps the same. However, in narrative videos, to help viewers better understand the plot, it would be useful to suggest the importance of each character. In the future, we can explore how to suggest the importance of characters via PIPs' size and appearance. For example, the size of PIPs can be different according to the character's importance. The color and thickness of the PIP border can also be adjusted to represent the importance of each character. The effects of bringing in such visual cues for importance suggestions to PIPs will need to be investigated.

9. Conclusions

This paper presents AdaPIP, an intuitive picture-in-picture view guiding method with adaptive view ranges and window sizes. To enhance viewers' watching experience, we conducted a study and formulated a content-related principle for adaptively adjusting the view range of the PIP planes. We also adapt our method and Outside-In in

an HMD-based VR environment engaged with controller-based interaction. Our method has shown statistical superiority over Outside-In in many aspects through a series of experiments in both 2D screen and VR environments. We will explore automatic labeling and tracking for future studies and how to assign different importance to PIPs.

Acknowledgement

This work was supported by the Natural Science Foundation of China (Project Number 62132012), Beijing Science and Technology Program (Project Number Z221100007722001), and Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] Adam Cosco. *Knives*, June 2019. 2, 3, 9, 12, 13

[2] M. Adcock, D. Feng, and B. Thomas. Visualization of off-surface 3d viewpoint locations in spatial augmented reality. In *Proceedings of the 1st symposium on Spatial user interaction*, pages 1–8, 2013. 3

[3] AutoNavi Information Technology Co.Ltd. *AutoNavi*, August 2021. 2, 5

[4] P. Baudisch and R. Rosenholtz. Halo: a technique for visualizing off-screen objects. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 481–488, 2003. 1, 3

[5] M. V. d. Broeck, F. Kawsar, and J. Schöning. It's all around you: Exploring 360 video viewing experiences on mobile devices. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 762–768, 2017. 4

[6] A. Cajar, R. Engbert, and J. Laubrock. Spatial frequency processing in the central and peripheral visual field during scene viewing. *Vision Research*, 127:186–197, 2016. 4

[7] Corridor. *360 Wizard Battle*, November 2016. 6

[8] E. David, J. Beitner, and M. L.-H. Võ. Effects of transient loss of vision on head and eye movements during visual search in a virtual environment. *Brain sciences*, 10(11):841, 2020. 4, 8

[9] E. J. David, P. Lebranchu, M. P. Da Silva, and P. Le Callet. Predicting artificial visual field losses: A gaze-based inference study. *Journal of vision*, 19(14):22–22, 2019. 4

[10] D. Fonseca and M. Kraus. A comparison of head-mounted and hand-held displays for 360 videos with focus on attitude and behavior change. In *Proceedings of the 20th International Academic Mindtrek Conference*, pages 287–296, 2016. 4

[11] Google LLC. *Google Map*, 2021. 2, 5

[12] Google Spotlight Stories. *Google Spotlight Stories: Special Delivery Trailer*, December 2015. 6

[13] Google Spotlight Stories. *360 Google Spotlight Stories: HELP*, April 2016. 6, 9, 12

[14] Google Spotlight Stories. *360 Google Spotlight Stories: Rain or Shine*, November 2016. 2, 3, 6

[15] Google Spotlight Stories. *360 Google Doodles/Spotlight Stories: Back to the Moon*, May 2018. 7, 8, 9, 12

[16] S. Gustafson, P. Baudisch, C. Gutwin, and P. Irani. Wedge: clutter-free visualization of off-screen locations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 787–796, 2008. 1, 3

[17] S. G. Gustafson and P. P. Irani. Comparing visualizations for tracking off-screen moving targets. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, pages 2399–2404, 2007. 1, 3, 4

[18] iNFINITE Production. *Crowd-Sourced Data*, June 2020. 4, 8, 10

[19] Iris. *Invisible - Episode 5 - Into The Den*, November 2016. 6

[20] S. Kasahara and J. Rekimoto. Jackin: integrating first-person view with out-of-body vision generation for human-human augmentation. In *Proceedings of the 5th augmented human international conference*, pages 1–8, 2014. 4, 8

[21] D. Kit, L. Katz, B. Sullivan, K. Snyder, D. Ballard, and M. Hayhoe. Eye movements, visual search and scene memory, in an immersive virtual environment. *PLoS One*, 9(4):e94362, 2014. 4, 8

[22] A. M. Larson and L. C. Loschky. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*, 9(10):6–6, 2009. 4

[23] C.-L. Li, M. P. Aivar, D. M. Kit, M. H. Tong, and M. M. Hayhoe. Memory and visual search in naturalistic 2d and 3d environments. *Journal of vision*, 16(8):9–9, 2016. 4

[24] Y.-C. Lin, Y.-J. Chang, H.-N. Hu, H.-T. Cheng, C.-W. Huang, and M. Sun. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2535–2545, 2017. 1, 3

[25] Y.-T. Lin, Y.-C. Liao, S.-Y. Teng, Y.-J. Chung, L. Chan, and B.-Y. Chen. Outside-in: Visualizing out-of-sight regions-of-interest in a 360 video using spatial picture-in-picture previews. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 255–265, 2017. 1, 3, 5, 8

[26] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2010. 13

[27] S. J. Liu, M. Agrawala, S. DiVerdi, and A. Hertzmann. View-dependent video textures for 360° video. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pages 249–262, 2019. 1, 3

[28] S. Matsuzoe, S. Jiang, M. Ueki, and K. Okabayashi. Intuitive visualization method for locating off-screen objects inspired by motion perception in peripheral vision. In *Proceedings of the 8th Augmented Human International Conference*, pages 1–4, 2017. 4, 8

[29] M. Millodot. *Dictionary of Optometry and Visual Science E-Book*. Elsevier Health Sciences, 2014. 4

[30] National Geographic. *Lions 360°*, June 2017. 9

[31] A. Nuthmann. On the visual span during object search in real-world scenes. *Visual Cognition*, 21(7):803–837, 2013. 4

[32] A. Pavel, B. Hartmann, and M. Agrawala. Shot orientation controls for interactive cinematography with 360 video. In

Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, pages 289–297, 2017. 1, 3

[33] T. Rhee, L. Petikam, B. Allen, and A. Chalmers. Mr360: Mixed reality rendering for 360 panoramic videos. *IEEE transactions on visualization and computer graphics (IEEE VR 2017)*, 23(4):1379–1388, 2017. 1

[34] The Rock. *The Rock Presents: "Escape From Calypso Island" - A 360 VR Adventure*, November 2016. 6

[35] S. Rothe, D. Buschek, and H. Hußmann. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction*, 3(1):19, 2019. 3

[36] Y. Sato, Y. Sugano, A. Sugimoto, Y. Kuno, and H. Koike. Sensing and controlling human gaze in daily living space for human-harmonized information environments. In *Human-Harmonized Information Technology, Volume 1*, pages 199–237. Springer, 2016. 8

[37] W. J. Tam, L. B. Stelmach, and P. J. Corriveau. Psychovisual aspects of viewing stereoscopic video sequences. In *Stereoscopic Displays and Virtual Reality Systems V*, volume 3295, pages 226–235. International Society for Optics and Photonics, 1998. 8

[38] F. Zhou, S. Bing Kang, and M. F. Cohen. Time-mapping using space-time saliency. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3358–3365, 2014. 13