Computational Visual Media
https://doi.org/10.1007/s41095-023-0347-3 Vol. 10, No. 3, June 2024, 487–503

Research Article

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], and Song-Hai Zhang[1] (✉)

Abstract 360° videos enable viewers to watch freely from different directions but inevitably prevent them from perceiving all the helpful information. To mitigate this problem, picture-in-picture (PIP) guidance was proposed using preview windows to show regions of interest (ROIs) outside the current view range. We identify several drawbacks of this representation and propose a new method for 360° film watching called AdaPIP. AdaPIP enhances traditional PIP by adaptively arranging preview windows with changeable view ranges and sizes. In addition, AdaPIP incorporates the advantage of arrow-based guidance by presenting circular windows with arrows attached to them to help users locate the corresponding ROIs more efficiently. We also adapted AdaPIP and Outside-In to HMD-based immersive virtual reality environments to demonstrate 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 (PIP); virtual reality (VR); visual guidance

1 Department of Computer Science and Technology, Tsinghua University, Beijing, China, and Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China. E-mail: Y.-X. Li, liyixiao20@mails.tsinghua.edu.cn; G. Luo, lg22@mails.tsinghua.edu.cn; Y.-K. Xu, xuyike@xiaomi.com; S.-H. Zhang, shz@tsinghua.edu.cn.

2 Key Laboratory of Space Utilization, Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China. E-mail: heyu@csu.ac.cn.

3 Victoria University of Wellington, Wellington, New Zealand. E-mail: fanglue.zhang@vuw.ac.nz.

Manuscript received: 2023-02-18; accepted: 2023-03-31

1 Introduction

Panoramic videos, also known as 360° videos, enable filmmakers to produce dynamic scenes that support viewers to watch from different virtual perspectives. Because of its low capture and display costs, it is commonly used as immersive content for Virtual Reality (VR) and Mixed Reality (MR) applications [1], 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 while watching a 360° movie. Because it is an unavoidable problem that users cannot perceive all the information, attempts have been made to alleviate this issue by guiding users to watch eventful parts using visual indicators, redirecting view rotation or displaying off-screen content. Visual indicators visualize the direction to regions of interest (ROIs) using symbolic diagrams [2–5]. They effectively indicate where a target is but lack visual content information. Navigation-based methods [2, 6, 7] change viewpoints automatically (auto-pilot) or inductively, forcing users to look in the direction of an important event. This method enables users to go through a series of events yet inevitably degrades immersion and prevents users from seeing multiple ROIs. The method relying on extra contents in Outside-In [8], 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.

In a pioneering study, Outside-In proposed the use of a set of 2D PIP windows to display off-screen ROIs. Although Outside-In has been demonstrated to outperform conventional arrow-based navigation

methods, it has the following drawbacks that limit 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 the PIPs remains constant, they often occlude and overlap. (2) Important content may not be significantly present in PIPs, especially when the target object is too close to or too far from the PIP's camera because its view range is constant. Some examples are presented in Fig. 1.

This paper proposes a PIP method with more accurate recommended content and optimized presentation. Here, we focus on character-based 360° videos, in which the ROI can be more clearly defined. For other types of videos, such as scenery videos, users may be interested in exploring the entire scene, making it difficult to define ROIs and the view direction guidance unnecessary. When playing character-based videos, making the audience focus on ROIs containing the characters' actions is essential for maintaining the narrative drive. To better attract users to the ROIs, an appropriate PIP method must display useful contextual information for viewers to understand the content in the guidance window. Different levels of significance and view ranges of the preview windows are required to guide users toward different characters. The direction indicator of a PIP window also considerably influences the effectiveness of view guidance. Conventional approaches, such as arrow-based guidance, can be considered for use in the PIP representation for ROI navigation, as they incur no additional learning costs, as shown in Fig. 2.

(a) Outside-In: distant character (b) AdaPIP: distant character

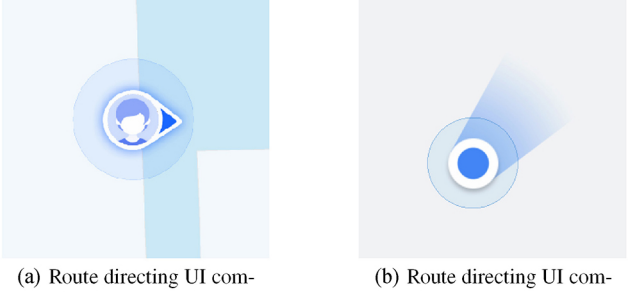
(c) Outside-In: close character (d) AdaPIP: close character

Fig. 1 Video play with Outside-In and AdaPIP. A delivery man appeared from the corner at the beginning of the video [9]. (a) Outside-In displays this event through a fix-range preview window, which is hard to notice. (b) AdaPIP adaptively reduces the context range of PIPs to visualize his movements clearly. (c, d) In the video [10], 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 on the above considerations, we present a new PIP-based guidance method that provides a better 360° film-watching experience, namely AdaPIP. Our method mitigates the aforementioned issues of the 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 [11]. Each off-screen character was 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 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 important content is infrequent. To determine the optimal context range displayed in PIPs and accurately present the off-screen characters, we conducted user studies and demonstrated 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 developed an adaptive PIP method in which the view range and included characters of a PIP window can be



(a) Route directing UI combination in AutoNavi (b) Route directing UI combination in Google Map

Fig. 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 or the isosceles trapezoid used in Google Map point to the target direction.

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 conducted extensive experiments to demonstrate the effectiveness of our system on both 2D screens and VR environments by comparing AdaPIP with Outside-In and a baseline method in which no directing technique was applied. Subjective ratings of several 360° video clips indicated the superiority of our method over Outside-In and the baseline method, demonstrating that AdaPIP provides a more comfortable and effective watching experience with a comparable degree of immersion. An additional test was conducted in both 2D and VR environments to demonstrate the benefits obtained by leveraging the adaptive mechanism for the content display of PIPs.

The contributions of this study 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 follows. Section 3 introduces the design of each element of AdaPIP. An adaptive scheme for dealing with different types of content is described and validated in Section 4. Section 5 details how we adapt AdaPIP and Outside-In to the VR environment. Section 6 explains the experiments to evaluate AdaPIP. Section 7 reports and analyzes the evaluation experiment results. Section 7.2 summarizes the feedback from the participants.

2 Related works

2.1 Attention guidance in virtual environment

Considerable research has been conducted to explore attention guidance techniques in AR and VR environments, as well as the particular case of a virtual environment: 360° video. Rothe et al. [12] divided visual guidance techniques into two categories: on-screen guidance and off-screen guidance. On-screen guidance focuses on guiding user fixation. By applying special screen effects (e.g., saliency modulation/blurring/stylistic rendering/gaze direction), users can be guided to focus on specific parts of a screen. However, this type of instruction can only be seen when it is inside the viewer's current field of view. Therefore, on-screen guidance has significant limitations when users can freely choose where to look. In contrast, off-screen guidance is dedicated to presenting off-screen content within the viewer's view range. Thus, we focused on the off-screen guidance technique in our study.

A popular off-screen guiding method uses graphics and symbolic figures to indicate out-of-view targets. For example, arrows [2], haloes [3], radar points [4], and wedges [5] are adopted to provide spatial clues. For virtual environments that allow free movement, Adcock et al. [13] proposed a composite wedge, 3D vector pairs, and a novel idea of rendering lit and shadowed areas to visualize the precise location of off-surface viewpoints in 3D space, thereby facilitating remote collaboration. By contrast, instead of using graphics, in a recent study, Outside-In creatively introduced a picture-in-picture guidance method. It directly presents ROIs in small inline windows that overlap with the main screen. More details about Outside-In will be discussed later in this section.

In addition, the force rotation method is effective in ensuring that users catch all important events. This method rotates the scene until the ROI is within the viewer's FOV. For example, Autopilot [2] 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 is used to provide viewers with important details. Pavel et al. [6] extended this experience to 360° videos by delivering important areas to viewers at every cut. However, whether the direction changes in the exact location due to the cuts cause disorientation still needs to be investigated [12]. In a recent work, Liu et al. [7] proposed a view-related playback method. They defined several gaze conditions (e.g., looking at a specific ROI) and seamlessly looped the gate clips until the conditions were met. Thus, viewers must turn to the given positions to see an important event. However, looped

audio introduces significant artifacts that significantly degrade the 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 view the content that is out of view and allows them to decide whether to look at it. One disadvantage of this method is that the inline windows always overlap on the main screen; thus, important content can be blocked out. Another disadvantage is the missing information regarding the position of the ROIs [12], which has been solved in Outside-In using the inline window itself as an arrow. Inspired by the concept of perspective projection, Lin et al. placed inline windows on the side near the ROI. They reshaped them according to their relative positions, making them appear to have the correct perspective relationship. Thus, users can naturally infer the positions of ROIs based on the appearance of the PIP planes.

However, the PIP windows of Outside-In inevitably obscure objects in the main window. To mitigate the occlusion problem, Lin et al. attempted to strike a balance by adjusting the size of the PIP plane based on the importance of the content behind it. Moreover, inline windows show only a fixed view range, which does not fit different situations very well. For example, as illustrated in Fig. 1, at the beginning of the video [9], a delivery man appeared from the corner, which is hard to notice in the PIP; in the video [10], a man moves towards the door and is very close to the camera, making the PIP only show his upper body. In addition, when two off-screen targets were close to each other, the PIP representing the farther target covered a large part of the PIP of the closer target. We address this issue by adaptively presenting the content and using a different layout.

2.3 Watching experience of head-mounted displays

Compared with a 2D screen, a VR headset such as an HMD provides a more immersive experience while watching 360° videos [14].

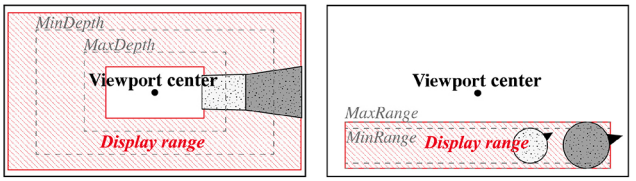
The heightened sense of immersion not only enriches the user's perception of presence but also elicits a stronger emotional response to visually appealing content [15]. However, this advantage comes at the cost of increased symptoms of nausea, oculomotor, and disorientation as illustrated in previous studies [4, 14].

VR headsets provide an authentic experience via the increment in both the horizontal and vertical FOVs (≈ 80°–174° for a horizontal FOV and ≈ 84°–114° for a vertical FOV [16]) compared to a 2D screen.

Drawing from the domain of visual perception, human vision can be divided into three principal regions: the fovea, parafovea, and periphery. The fovea constitutes the central 2° of vision, whereas the parafovea encompasses a circumference of approximately 5° from the point of fixation. Collectively, these two regions are commonly referred to as central vision. Beyond the parafoveal area lies the peripheral region, commonly referred to as peripheral vision [17]. Central vision is responsible for perceiving high vision, shapes, and colors. However, peripheral vision is not sufficiently accurate to perceive highly diverse visual content and is used to target the next eye movement [18]. The wider FOV of a VR headset also provides users with more peripheral vision than a 2D tablet screen (e.g., VR headsets provide ≈ 90 degrees peripheral vision) [19–21] and 2D screens ≈ 30 × 20 degrees [22–24]). This created sufficient room for displaying guide elements. Occlusion and interference issues can be mitigated by placing guide elements in peripheral areas while maintaining the ability to guide the directions [25, 26].

3 Design of AdaPIP

The basic layout of AdaPIPs is illustrated in Fig. 3. We displayed the off-screen targets in 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



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

Fig. 3 A comparison of the display range between Outside-In and our method.

user continuously rotates their views or the off-screen target moves. In addition, an arrow was attached to the 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 operate 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 contained the original 360° video along with several annotation files, including (1) manually specified characters and (2) tracking data of the characters (both spatial and temporal). The existing video tracking algorithm is not sufficiently robust to provide sufficiently accurate object masks for 360° videos, particularly when the videos contain cartoon characters or have poor lighting conditions. Because our work does not aim 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 certain keyframes and obtain their motion paths via piecewise linear interpolation.

To reduce the occlusion caused by PIPs, we chose to render preview windows on the lower part of the view window. While watching a video on a 2D display, the user's sight is typically perpendicular to the screen. Thus, we superimpose the PIP image planes onto a 360° video. In a VR environment, the 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 m to support possible real-time interactions because this distance can be easily reached in a VR environment.

In the following subsections, we first introduce how to determine the sizes and positions of a PIP preview window and its arrow to indicate the distance and direction of an object intuitively. We then 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 method to indicate the distance between the target object and the user's current viewpoint. Thus, we made the PIP windows slide horizontally in the specified region when the user or off-screen target moved. Given a 360° video represented by equirectangular projection, we use latitude and longitude to define the unique position on a frame, where the 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 using the normalized Euclidean distance, which has a value between 0 and 1:

$$D = \sqrt{\left[\left(\frac{\Delta\text{latitude}}{90^\circ}\right)^2 + \left(\frac{\Delta\text{longitude}}{180^\circ}\right)^2\right]/2} \quad (1)$$

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

Apart from the position of the PIP, its attached arrow also indicates the relative distance as 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 appears to stretch. 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 determined by the distance D and the predefined maximum/minimum arrow lengths $L_{\max/\min}$:

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

Please refer to our video in the Electronic Supplementary Material (ESM) for how the PIPs work when users watch 360° videos.

3.2 Direction representation

To represent the direction of rotation of the target object, we rotated 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 type of indication method is often used in navigation applications and has been shown to be effective in reducing the learning costs for users. In addition, because we use the arrow instead of the PIP itself to

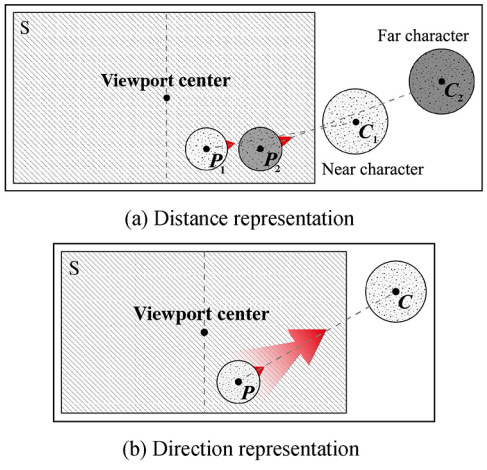


Fig. 4 Direction and distance representation in AdaPIP. (a) The distances between the enclosed characters and the user are represented by the distance the PIP deviates from the viewport center. In other words, when the user approaches the character outside the screen, the PIP will be closer to the viewport 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.

indicate the location of the target, as in Outside-In, our PIP window no longer requires a large display area, which can mitigate the occlusion issue. See Fig. 3.

3.3 View-dependent interactions

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

Display visibility. Our interaction scheme operates when the user starts watching a 360° film. If the user is not looking at a specific character, the PIP for that character appears. After the user turns their head toward the character, the PIP fades out. In addition, we implemented auto-pilot interactions. By clicking on any PIP, the user's view can be directly turned in the corresponding direction of 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 used the distance to the nearest target to determine the window size S of all the PIPs by

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

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

4 Adaptive context

In previous studies, such as Outside-In, an off-screen ROI on the PIP plane was rendered with the same FOV as the main window. This causes serious issues when the characters inside an off-screen ROI are too far or too close to the viewpoint. The characters may be too small to be observed when they are far away from the user's viewpoint and too large to be displayed entirely when they are close to the viewpoint. Because the PIP windows should focus on the characters rather than the entire ROI area, 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 conducted to reveal users' preferences for the view range of PIPs. Therefore, 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 prefer the view range of the content in PIPs? More specifically, do users prefer a wider range with more context or a narrower range?

RQ2: When multiple characters are 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 six character-based 360° videos from YouTube [9, 27–31]. We then extracted eight video clips of 10–15 s from the above six videos according to the size of the characters and the relative distance between the characters. These eight video clips were divided into two groups of four according to the size of the characters: large-character video clips (LC) and small-character video clips (SC). In the LC video group, there were two video clips (LCF) with two characters far apart and two video clips (LCC) with two characters close together, similar to the SC video group (SCF and SCC). See Fig. 5. We used video

Fig. 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 Refs. [27–29, 31].

clips with lengths of 10–15 s 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 used a circular bounding box to represent the character and considered the center of mass of the circle as the character center, as shown in Fig. 6. The character's moving path was recorded as the path of its bounding box via a semi-manual annotation process introduced in Section 2.

We designed two experiments to answer the aforementioned research questions: In the experiment for RQ1, each participant watched two LCF videos and two SCF videos. We randomly assigned narrow PIPs and wide PIPs to the played videos, where a narrow PIP displayed a 10% larger area than the character's bounding box and a wide PIP displayed a 60% larger area. In the experiment for RQ2, we provided two LCC and two 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 in a single PIP window; for separate PIPs, we assign a PIP window for each character.

Fig. 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 Ref. [32].

4.1.2 Procedure and measures

We recruited 10 participants (six males and four females) for the two experiments. The participants were all college students aged 19–28, and seven of them had previously watched 360° videos on 2D screens. We designed two experiments for the two research questions accordingly. At the beginning of each experiment, a brief tutorial was provided to the participants explaining the PIP-based guidance method used in the experiment (wide-range and narrow-range PIPs for the first experiment, PIPs with grouped characters, and separate characters for the second experiment). The participants were encouraged to try different view directions to understand how AdaPIPs worked when watching the test videos. After the tutorial, the participants were asked to watch 360° videos using different PIP-based guidance methods. They were informed which video type (LC or SC) and which method would be shown before watching. For each experiment, the participants were required to watch four video clips; therefore, each participant needed to watch $2 \times 4 = 8$ video clips throughout the study. The order in which the videos were played was randomized for different participants. After watching a group of videos, the participants were asked to rate their watching experience using a score between 1 (worst) and 7 (best).

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

4.1.3 Results

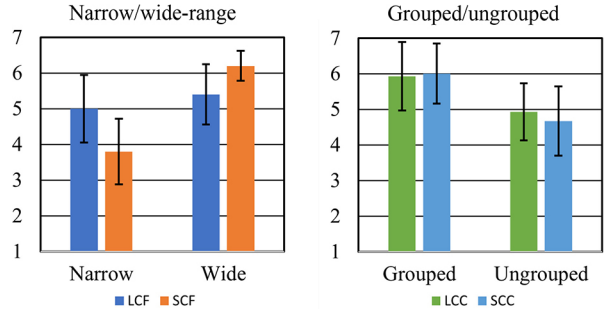
Context range. In experiment 1, we collected 10 (participants) × 2 (methods) × 2 (video types) = 40 ratings. For videos with large characters, the

ratings of narrow-range PIPs ($\mu = 5$, $\sigma = 0.943$) and large-range PIPs ($\mu = 5.4$, $\sigma = 0.843$) did not differ significantly. For small-character videos, we found that the participants preferred a large range ($\mu = 6.2$, $\sigma = 0.422$) to a narrow range ($\mu = 3.8$, $\sigma = 0.919$). See Table 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 ($p = 0.686$). Pairwise comparisons also showed a significant difference between the ranges for the small-character videos. This demonstrates that users have no obvious preference for the display range for large characters but prefer a wide context range for small characters.

Grouped or separate characters. The same as in experiment 1, we got another 40 sets of ratings in experiment 2. We found that all the participants gave the grouped characters higher scores. Some expressed that grouping characters creates fewer distractions and helps them see the interactions among characters, encouraging them to watch the corresponding event using their main-view window. As indicated in Table 1, for large-character videos, using PIPs with grouped characters ($\mu = 6.1$, $\sigma = 0.568$) got a higher score than ungrouped ($\mu = 3.7$, $\sigma = 0.675$), the same for the small-character videos (grouped: $\mu = 6.2$, $\sigma = 0.422$, ungrouped: $\mu = 2.6$, $\sigma = 0.843$). We further analyzed the results using a two-way repeated-measures ANOVA, where a statistically significant interaction between the methods and videos was found, with $F(1, 9) = 36$ and $p < 0.05$. Therefore, the effects of both the grouped and ungrouped methods were analyzed in each video, and a significant effect was found for both large- and small-character videos. We also analyzed the effects of the videos on each method and only found a significant effect for the ungrouped method. 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 rated the ungrouped method worse for the SC videos.

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



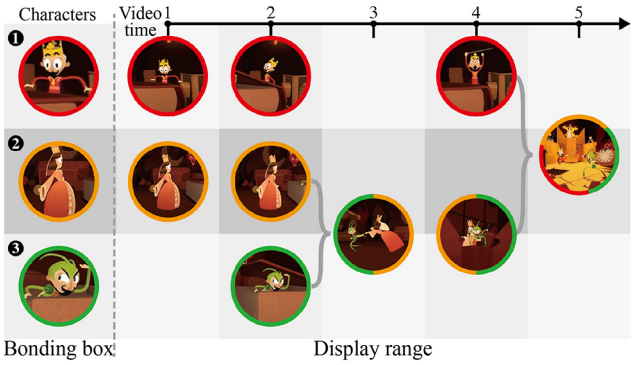


Fig. 7 Illustration of the adaptive context mechanism. Smaller characters have more contextual information rendered on the PIPs; larger characters have less contextual information; when the characters' bounding boxes intersect, we consider these characters to have possible interactions in that frame 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 Ref. [32].

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 the characters. Therefore, we applied a three-stage process to dynamically adjust the position and the context range of PIPs based on the experimental results.

First, we counted the number of interactions among characters and grouped the characters in different time ranges because users clearly prefer whether the characters should be grouped when they have close relationships. When the bounding boxes of characters intersect, we consider them to have possible interactions in that frame. If the duration of the intersection exceeds a specified threshold ($\eta = 10$), 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 they still exist in the picture, ensuring the temporal stability of the relationship. 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 did not interact with others, it formed a group.

Second, we calculated the center position and context range of PIPs based on the bounding boxes of the character groups in each frame. We used a wide range for a group with a single small-size character. If multiple characters were present in one group, we performed a weighted average to calculate the center.

Finally, we checked how the groups changed over time and made smooth transitions between the different context ranges. For example, before and after the merging or separation of groups, there are noticeable context ranges and character size changes. To ensure the smoothness of these changes, the context range of the transition frames was interpolated by Laplace Smoothing using 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 to using only a narrow context range, we provide the necessary background and interaction information. We also enabled an AdaPIP window to dynamically adapt to the size changes of the characters. When a character moves towards the camera, the view range increases such that users can notice the change in the distance of the character.

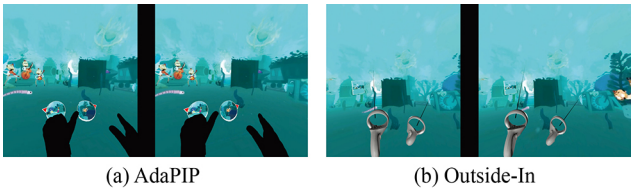
5 Adaption to VR environment

For a more comprehensive assessment of the AdaPIP method, we also explored how to adapt PIP technologies to 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 [16, 19, 21, 33]. The peripheral area can be used to effectively display guide elements while mitigating occlusion and interference issues [25, 26], indicating an enlarged display area for prompt windows when adapting Outside-In and AdaPIP to VR environments. Moreover, both Outside-In and AdaPIP superimpose prompt windows on the original 360° video when it is played on 2D screens. To inherit this idea, we placed each PIP on a plane at different depths from a 360° video. We have also enabled a stereoscopic display for a higher sense of perceived depth [34].

AdaPIP. In a 2D scenario, users can create an autopilot in the viewing direction for an ROI by clicking on 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. In addition, the pitch rotation of the viewing direction in a VR environment may confuse the user's navigation. For example, when a user creates an autopilot in the sky, their 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 limited the pitch rotation and only allowed yaw rotation when an autopilot was used in the VR. We indicated the pitch angle to the ROI center using a stretched arrow after autopilot.

Outside-In. Outside-In places the PIP windows around the screen center in 2D [8]. In the VR environment, we adopt the same method and limit the display range of the PIP windows to the peripheral area. Unlike AdaPIP, the depth of each PIP plane in Outside-In implies the distance to the corresponding object, which linearly decreases with the distance between the target object and the view center.

We used a large depth range to achieve an appearance similar to that of the original 2D outside-in. However, this causes the PIP plane to be too far from the user to reach the autopilot. To solve this issue, a virtual ray emitted from the controller was used to hit the PIP plane, and the user could press the controller button to trigger autopilot (see Fig. 8). To avoid the aforementioned navigation confusion, we also limited the pitch rotation and used the PIP's position to indicate the required pitch rotation after autopiloting.



(a) AdaPIP (b) Outside-In

Fig. 8 Binocular view of the user interface of AdaPIP and Outside-In in a 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 Ref. [32].

6 User experiment settings

To test whether our approach improves user experience, we conducted 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 four videos covering a variety of genres from YouTube to demonstrate the generalizability of our method for different narrative types. Detailed information is provided in Table 2. Because our method focuses on characters in videos, we did not use scenery videos. We divided each video into two discontinuous video clips with durations ranging from 49 to 70 s and randomly labeled each video clip as 1 or 2, as clip 1 was always displayed in a 2D environment and clip 2 in a VR environment.

Table 2 Summary of example videos

Video name	Genre	Author	Length (s)	
			Clip 1	Clip 2
Back to the moon [32]	Comedy	Google Spotlight	51	66
Help [27]	Horror	Google Spotlight	53	49
Knives [10]	Thriller	Indie	57	51
Lions [35]	Documentary	National Geographic	66	70

6.1 Method

To test the effectiveness of our method in 2D and VR environments, our user study included two formal tests: a **2D screen test** and a **VR test**. The participants were asked to take one of these two tests. Before the formal test, participants were given a brief introduction to all the 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 became familiar with the different methods. After the formal test, an **extra test** was conducted to further explore whether AdaPIP could help the participants recognize the prompt content.

2D screen test. For this test, participants were asked to watch four video clips three times on a desktop monitor and were informed of the PIP technique. One of the following 3 methods was applied 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 counterbalanced, as was the order of the video clips presented to the participants. The participants were asked to fill out a questionnaire after watching one video using different PIP methods. They were also asked to rate three methods in terms of Q1: overall performance and Q2: understanding the level of spatial relationship. They were also asked to rate Outside-In and AdaPIP in terms of Q3: interference level and Q4: recognition level of the prompt content.

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

Fig. 9 Setup for (a) 2D screen test and (b) VR test.

VR test. For the VR test, the participants were required to wear an HMD with two controllers. They were asked to watch four video clips three times using different PIP techniques. The order of the three methods and video clips was the same as in the 2D test. After watching the video three times, the participants removed the HMD, took a break while filling out a questionnaire, and rated their experience using the same criteria as the 2D test.

Extra test. We conducted an extra test to validate whether the adaptive mechanism could 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 four videos they saw in the 2D screen test through a desktop monitor, and participants who took the VR test were asked to watch the same four videos as the VR test wearing an HMD. All participants were asked to watch each video twice, applying AdaPIP or its nonadaptive version in random order. The order of the video clips was randomized and counterbalanced. After watching the

video twice, participants were required to rate the recognition level of the prompt content for the two PIP methods.

6.2 Participants

We recruited 28 (16 males and 12 females) university students with different majors, aged 18–26 years old, as participants. 13 (eight males and five females) of them signed up for the 2D screen test, and the remaining 15 (eight males and seven females) signed up for the VR test. For participants taking the 2D screen test, seven had watched a 360° video via 2D screen previously; for participants taking the VR test, none had worn an HMD to view a 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 was approximately 40 cm. We built our platform in Unity (Version 2019.4.22F1C1) and played 360° videos using Unity's Play Window (1778 pixels×1000 pixels). Participants could 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. Oculus Quest 2 has a single-eye resolution of 1832 pixels×1920 pixels with a horizontal FOV of 89 degrees (±4°) and a vertical FOV of 93 degrees (±5.1°) [16]. We played 360 videos on the Unity platform and streamed them to Oculus Quest 2. A swivel chair was provided to the participants, which was able to rotate easily in the yaw dimension. The participants were asked to use the two controllers while watching the video. They could jump to the corresponding view by touching the PIP window via the controllers.

6.4 Measurements

After viewing each video clip, the 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 on how the participants assessed the assistance of each technique and the aspects they liked or disliked. The full 2D screen test, including practice, interviews, and the extra test, lasted approximately 40–50 min, while the entire VR test lasted approximately 50–60 min.

7 Results

7.1 Subjective rating

7.1.1 2D screen test

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

Q1: Overall performance. Based on the results presented in Table 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 the different methods, $F(2,24) = 17.334$, $p < 0.0001$. No statistically significant interaction was found between the different videos or between the methods and videos. Pairwise t-test comparisons demonstrated significant differences between the methods.

Q2: Understanding level of spatial relationship. As shown in Table 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 the different methods was found, with $F(2,24) = 54.194$ and $p < 0.0001$. No statistically significant interaction was found between the different videos or between the 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 Table 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 the different methods was found, with $F(1,12) = 39.103$ and $p < 0.01$. No statistically significant interaction was found between the different videos or between the methods and videos.

Q4: Recognition level of the prompt content. AdaPIP and Outside-In showed similar levels of readability for the content prompted by the PIPs; see Table 3. We did not find significant interactions among the methods, videos, methods, and videos.

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 lower 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	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

7.1.2 VR test

For the VR test, we collected ratings from 15 participants based on the aforementioned criteria.

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

Q2: Understanding level of spatial relationship. As shown in Table 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 the different methods was found, with $F(2, 28) = 24.702$ and $p < 0.0001$. No statistically significant interactions were found between the different videos, and no statistically significant interactions were found between the methods and videos.

Q3: Interference level. Similar to the 2D screen test, higher scores represented higher interference to evaluate the interference level. As shown in Table 4, most participants rated Outside-In higher, feeling that Outside-In caused more distractions when watching videos. No statistically significant interaction was found between the different videos or between the methods and videos. We also performed paired t-test comparisons, which showed that the scores of the different methods were significantly different.

Q4: Recognition level of the prompt content. As revealed in Table 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 that both of them could effectively help them understand the plot. A two-way repeated-measures ANOVA was performed and no significant interaction was found between the different methods, between the different videos, or between the method and video.

7.1.3 Extra test

2D screen environment. We collected the ratings from 13 participants for the extra test in a 2D environment. Based on the results presented in Table 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 assigned higher scores to the former, see Fig. 10.

A two-way repeated-measures ANOVA was performed, and a statistically significant interaction

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 lower 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

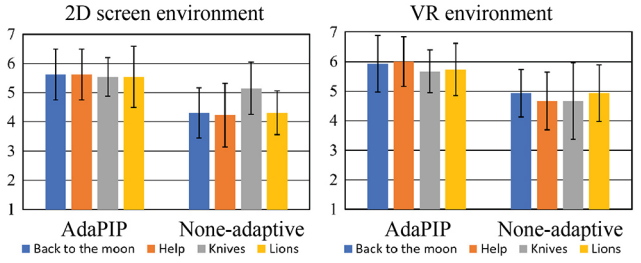


Fig. 10 Average ratings for the extra test in 2D screen environment (left) and VR environment (right). Error bars show standard deviations.

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 for each video. *P*-values were adjusted using the Bonferroni multiple-testing correction method. The effect of the treatment was significant for the video Back to the moon, Help, Lion but not for the video Knives. Pairwise comparisons, using paired t-tests, showed 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 nonadaptive version for video Knives. However, for the other three videos, a statistically significant difference was observed in the scores between the two methods. By inspecting the video "Knives" [10], we found that its character sizes are moderate and keep nearly constant. Therefore, AdaPIP's results are similar to those of the non-adaptive PIP methods only in this video.

VR environment. We collected the 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 Table 5 and Fig. 10, AdaPIP outperforms the non-adaptive method. We further performed a two-way repeated-measures ANOVA and found a statistically significant interaction between the different methods, with $F(1, 14) = 75.162$ and $p < 0.0001$. There were no statistically significant interactions between the different videos or between the methods and videos.

7.2 Interviews

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

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

	Back to the moon				Help			
	AdaPIP		None-adaptive		AdaPIP		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				**Lion**			
	AdaPIP		None-adaptive		AdaPIP		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

7.2.1 Overall preference

In a 2D environment, 10 participants felt that adding PIP windows to 360° videos could enhance their watching experience, whereas three other participants preferred 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 four participants expressed their preference for Outside-In. "Outside-In feels like surveillance windows." P6 said, "With these windows, I can monitor every event in all directions."

While watching with VR headsets, 13 participants claimed a preference for watching with guidance. Two participants believed that the necessity of adding PIP windows depended on the video content. For videos, such as Back to the moon [32], changes in light and scenery can effectively guide participants to focus on the protagonist, and it is not necessary to add extra guidance. However, for videos such as Help [27], the protagonists constantly move from side to side at high speeds throughout the video; in this case, PIP windows are needed. In addition, eight participants said that, despite having a swivel chair, they still disliked turning their heads or bodies while watching the video. It enhanced their viewing experience if they could see any plot without turning their heads or bodies. Thirteen participants expressed their preference for AdaPIP; one participant said he had no particular preference; P18 said he preferred Outside-In because "Outside-In displays out-of-view content in a larger window and is easier to recognize."

7.2.2 Spatial guidance

Of the 28 participants, 26 said that AdaPIP was more effective in guiding direction. Five participants stated 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 is why I think Outside-In performs better."

7.2.3 Context range

Among the 28 participants, eight 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; two participants said AdaPIP's PIP window could display clearer content. In addition, six participants expressed a 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, five participants mentioned the size of the characters in the PIP windows. P26 said, "It is 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 than with Outside-In. The two participants believed that 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 is distracting while watching the video."

7.2.5 Interactions in VR

Ten of the fifteen participants preferred the interaction method of AdaPIP to trigger 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." Eight participants said that while this was a novel interaction, having to raise their hands every time for an 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 obtain a better viewing experience with the help of PIP guidance, and (2) our method is better evaluated than Outside-In. On the one hand, our method can effectively guide users to find targets and improve their understanding of spatial relationships. Moreover, 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 showed that our method can prompt more accurate content by adopting the adaptive context range. The only exception in our experiments was the video Knives [10], where the characters maintained a modest and constant size. For most videos, the proposed 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, the participants felt that the two types of interactions had both advantages and

disadvantages. Using virtual hands to touch PIPs in AdaPIP results in a lower level of interference and provides a novel experience; however, it requires frequent hand movements. Outside-In's ray-based interaction mitigates the issue of fatigue but causes more interference.

Overall, our method performed better 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 was designed to work with films of character-based stories. For videos captured for users to explore freely, such as scenery videos, the proposed method is not applicable. In addition, there are distraction issues. Some users mentioned 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, viewers may be less patient in watching the content of the PIP preview windows. However, we also believe that in most 360° videos, encouraging users to change their viewpoint can encourage them to fully explore the 360° virtual space.

8.2 Future works

Automatic labeling and tracking. In this study, characters were manually annotated for each video to ensure the accuracy of the PIP preview. The process requires labor-intensive work, especially for long videos. This step can be replaced by leveraging video segmentation and object-tracking methods [36, 37]. Because our algorithm requires almost negligible time, it 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 was always the same. However, in narrative videos, it would be useful to suggest the importance of each character to help viewers better understand the plot. In the future, we can explore how to suggest the importance of characters based on the size and appearance of PIPs. For example, the size of PIPs can be different according to the importance of the character. 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 presented 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 to adaptively adjust the view range of the PIP planes. We also adapted our method and Outside-In to an HMD-based VR environment engaged with controller-based Through a series of experiments in both 2D screens and VR environments, our method showed statistical superiority over Outside-In in many aspects. We will explore automatic labeling and tracking in future studies and how to assign different importance to PIPs.

Acknowledgements

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

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Electronic Supplementary Material

Electronic supplementary material is available in the online version of this article at `https://doi.org/10.1007/s41095-023-0347-3`.

References

[1] Rhee, T.; Petikam, L.; Allen, B.; Chalmers, A. MR360: Mixed reality rendering for 360° panoramic videos. *IEEE Transactions on Visualization and Computer Graphics* Vol. 23, No. 4, 1379–1388, 2017.

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

[3] Baudisch, P.; Rosenholtz, R. Halo: A technique for visualizing off-screen objects. In: Proceedings of the

SIGCHI Conference on Human Factors in Computing Systems, 481–488, 2003.

[4] Gustafson, S. G.; Irani, P. P. Comparing visualizations for tracking off-screen moving targets. In: Proceedings of the CHI '07 Extended Abstracts on Human Factors in Computing Systems, 2399–2404, 2007.

[5] Gustafson, S.; Baudisch, P.; Gutwin, C.; Irani, P. Wedge: Clutter-free visualization of off-screen locations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 787–796, 2008.

[6] Pavel, A.; Hartmann, B.; Agrawala, M. Shot orientation controls for interactive cinematography with 360 video. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, 289–297, 2017.

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

[8] Lin, Y. T.; Liao, Y. C.; Teng, S. Y.; Chung, Y. J.; Chan, L.; Chen, B. Y. 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, 255–265, 2017.

[9] Google Spotlight Stories. 360 Google Spotlight Stories: Rain or Shine. 2016. Available at https://www.youtube.com/watch?v=QXF7uGfopnY

[10] Adam Cosco. Knives. 2019. Available at https://youtu.be/IrAXKwEKVGA?si=y9gyhtBvxzFY1v-S

[11] AutoNavi Information Technology Co. Ltd. AutoNavi. 2021. Available at https://mobile.amap.com/

[12] Rothe, S.; Buschek, D.; Hußmann, H. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction* Vol. 3, No. 1, 19, 2019.

[13] Adcock, M.; Feng, D.; Thomas, B. Visualization of off-surface 3D viewpoint locations in spatial augmented reality. In: Proceedings of the 1st Symposium on Spatial User Interaction, 1–8, 2013.

[14] Van den Broeck, M.; Kawsar, F.; Schöning, J. It's all around you: Exploring 360° video viewing experiences on mobile devices. In: Proceedings of the 25th ACM International Conference on Multimedia, 762–768, 2017.

[15] Fonseca, D.; Kraus, M. 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, 287–296, 2016.

[16] iNFINITE Production. Crowd-Sourced Data. 2020. Available at https://www.infinite.cz/projects/HMD-tester-virtual-reality-headset-database-utility

[17] Larson, A. M.; Loschky, L. C. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision* Vol. 9, No. 10, 6.1–6.16, 2009.

[18] Millodot, M. *Dictionary of Optometry and Visual Science E-Book*. Butterworth-Heinemann, 2014.

[19] Kit, D.; Katz, L.; Sullivan, B.; Snyder, K.; Ballard, D.; Hayhoe, M. Eye movements, visual search and scene memory, in an immersive virtual environment. *PLoS One* Vol. 9, No. 4, e94362, 2014.

[20] Li, C. L.; Aivar, M. P.; Kit, D. M.; Tong, M. H.; Hayhoe, M. M. Memory and visual search in naturalistic 2D and 3D environments. *Journal of Vision* Vol. 16, No. 8, Article No. 9, 2016.

[21] David, E.; Beitner, J.; Võ, M. L. H. Effects of transient loss of vision on head and eye movements during visual search in a virtual environment. *Brain Sciences* Vol. 10, No. 11, Article No. 841, 2020.

[22] Nuthmann, A. On the visual span during object search in real-world scenes. *Visual Cognition* Vol. 21, No. 7, 803–837, 2013.

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

[24] David, E. J.; Lebranchu, P.; Perreira Da Silva, M.; Le Callet, P. Predicting artificial visual field losses: A gaze-based inference study. *Journal of Vision* Vol. 19, No. 14, Article No. 22, 2019.

[25] Matsuzoe, S.; Jiang, S.; Ueki, M.; Okabayashi, K. Intuitive visualization method for locating off-screen objects inspired by motion perception in peripheral vision. In: Proceedings of the 8th Augmented Human International Conference, Article No. 29, 2017.

[26] Kasahara, S.; Rekimoto, J. JackIn: Integrating first-person view with out-of-body vision generation for human-human augmentation. In: Proceedings of the 5th Augmented Human International Conference, Article No. 46, 2014.

[27] Google Spotlight Stories. 360 Google Spotlight Stories: HELP. 2016. Available at https://www.youtube.com/watch?v=G-XZhKqQAHU

[28] Corridor. 360 Wizard Battle. 2016. Available at https://youtu.be/bb5eETSspVI?si=Wayr9bbhRsVtrWSG

[29] Iris. Invisible - Episode 5 - Into The Den. 2016. Available at https://youtu.be/qYxNCB678WQ?si=uJhsaetH-HytKyzY

[30] The Rock. The Rock Presents: "Escape From Calypso Island" - A 360 VR Adventure. 2016. Available at https://youtu.be/G4w_MBMNMEQ?si=XGdQOCgb2-yy5XD8K

[31] Google Spotlight Stories. Google Spotlight Stories: Special Delivery Trailer. 2015. Available at https://youtu.be/3QxZtQoAIOs?si=Wz2pRXtEvRwLr5E6

[32] Google Spotlight Stories. 360 Google Doodles/Spotlight Stories: Back to the Moon. 2018. Available at https://youtu.be/BEePFpC9qG8?si=PxDQjkefXBOuUMd1

[33] Sato, Y.; Sugano, Y.; Sugimoto, A.; Kuno, Y.; Koike, H. Sensing and controlling human gaze in daily living space for human-harmonized information environments. In: *Human-Harmonized Information Technology, Volume 1*. Nishida, T. Ed. Springer Tokyo, 199–237, 2016.

[34] Tam, W. J.; Stelmach, L. B.; Corriveau, P. J. Psychovisual aspects of viewing stereoscopic video sequences. In: Proceedings of the SPIE 3295, Stereoscopic Displays and Virtual Reality Systems V, 226–235, 1998.

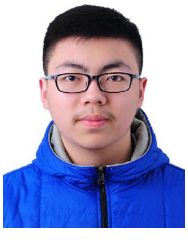
[35] National Geographic. Lions 360°. 2017. Available at https://youtu.be/sPyAQQklc1s?si=ztk3XKDkXchZqTCn

[36] Zhou, F.; Kang, S. B.; Cohen, M. F. Time-mapping using space-time saliency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3358–3365, 2014.

[37] Liu, C.; Yuen, J.; Torralba, A. SIFT flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 33, No. 5, 978–994, 2011.

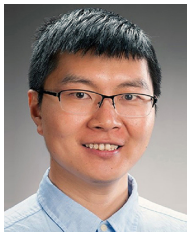
Yi-Xiao Li received her bachelor degree in arts & design from Tsinghua University, Beijing, in 2020, where she is currently pursuing her master degree in the Academy of Arts & Design of the same university. Her research interests include human–computer interaction and virtual reality.

Guan Luo is currently a Ph.D. student in the Department of Computer Science and Technology, Tsinghua University, supervised by Song-Hai Zhang. His research interests include computer vision and virtual reality.



Yi-Ke Xu is currently an undergraduate student in the Department of Computer Science and Technology, Tsinghua University. His research interests include virtual reality and image/video processing.

Yu He received a doctoral degree from the Zhejiang University of Technology in 2019. He completed his postdoctoral work in the Department of Computer Science and Technology at Tsinghua University in 2021. He is currently an assistant researcher at Yanqi Lake Beijing Institute of Mathematical Sciences and Applications. His research interests include 3D vision and virtual reality.



Fang-Lue Zhang is currently a lecturer at Victoria University of Wellington, New Zealand. He received his bachelor degree from Zhejiang University, Hangzhou, China, in 2009, and his doctoral degree from Tsinghua University, Beijing, China, in 2015. His research interests include image and video editing, computer vision, and computer graphics. He is a member of IEEE and ACM. He received Victoria Early Career Research Excellence Award in 2019.

Song-Hai Zhang received his Ph.D. degree in computer science and technology from Tsinghua University, Beijing, in 2007. He is currently an associate professor in the Department of Computer Science and Technology at Tsinghua University. His research interests include virtual reality and image/video processing.