Tourgether360: Collaborative Exploration of 360° Videos using Pseudo-Spatial Navigation

KARTIKAEYA KUMAR, Indian Institute of Technology at Guwahati, India
LEV PORETSKI, University of Toronto, Canada
JIANNAN LI, University of Toronto, Canada
ANTHONY TANG, University of Toronto, Canada

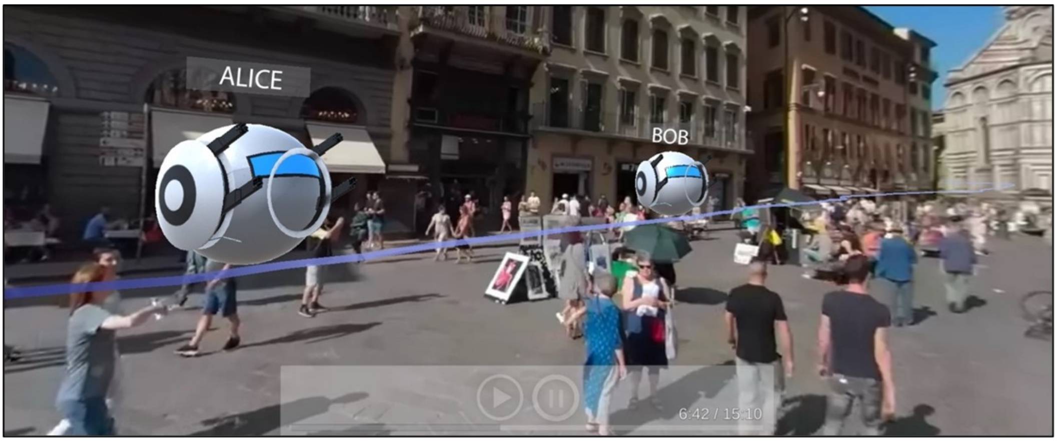

Fig. 1. Tourgether360 allows collaborators, represented by avatars, to tour together “inside” a 360-degree video.

Collaborative exploration of 360° videos with contemporary interfaces is challenging because collaborators do not have awareness of one another’s viewing activities. Tourgether360 enhances social exploration of 360° tour videos using a pseudo-spatial navigation technique that provides both an overhead “context” view of the environment as a minimap, as well as a shared pseudo-3D environment for exploring the video. Collaborators are embodied as avatars along a track depending on their position in the video timeline and can point and synchronize their playback. We evaluated the Tourgether360 concept through two studies: first, a comparative study with a simplified version of Tourgether360 with collaborator embodiments and a minimap versus a conventional interface; second, an exploratory study where we studied how collaborators used Tourgether360 to navigate and explore 360° environments together. We found that participants adopted the Tourgether360 approach with ease and enjoyed the shared social aspects of the experience. Participants reported finding the experience similar to an interactive social video game.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Conference acronym ’XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Kartikaeya Kumar, Lev Poretski, Jiannan Li, and Anthony Tang. 2018. Tourgether360: Collaborative Exploration of 360° Videos using Pseudo-Spatial Navigation. In . ACM, New York, NY, USA, 27 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

360° tour videos are an increasingly popular way of exploring remote destinations and environments. Such videos are typically shot using an omnidirectional camera mounted atop a tripod as the cameraperson moves through an environment (e.g., by walking or driving). The videos provide viewers with the ability to freely look around, independent of the direction that the cameraperson was moving. Because of this freedom, they provide users with a rich sense of immersion—particularly when coupled with head mounted displays (e.g. [6, 35]). Such videos are also increasing in popularity: beyond simply being used to view and compare vacation destinations, they are now increasingly being used by families to visit prospective college campus, or by realtors showing off homes for rent or sale. Thus, collaborative viewing is also becoming increasingly important. Collocated or remote friends may want to watch 360° videos to experience immersive and social entertainment together, or such videos may be used in the educational context, with a class of students going on a virtual museum tour or virtual trip to cultural locations.

The problem is that current 360° interfaces do not provide effective support for collaborative navigation and exploration of 360° videos (e.g. [34]). With only a handful of exceptions (e.g., [20, 22, 31] 360° video players are intended for single-person use; in addition to normal video playback controls, such video players need to provide a special, separate means for controlling the view orientation. On a desktop, orientation is controlled by grabbing the scene and moving it; on tablets, this is augmented through gyroscopic sensors, and on a head-mounted display, orientation is controlled by turning or tilting one's head. Yet, there is little to no support for collaborative viewing of these immersive videos. Prior work has demonstrated that when collaborators are watching 360° videos together, collaborators may not want to be looking at the same thing at the same time [33, 34]; in spite of this, they still want to maintain an awareness of what their collaborators are watching. While streaming 360° videos to remote users had been previously explored to some extent [20], we have yet to see collaborative 360° video viewing experiences that support these needs.

We propose a pseudo-spatial navigation metaphor for collaborative exploration of 360° videos inspired by a focus+context approach [32]. We realize this approach in a prototype system called Tourgether360, which allows several collaborators to explore a 360° video together. With Tourgether360, the video tour context is visualized as a path on a 2D map of the environment. Building on this approach first illustrated by Noronha and colleagues [24], users can scrub along the path to navigate both time and position in the 360 video (both in focus and in context views). This eases the coordination between finding points in the video with spatial locations visited in the video. We call this approach "pseudo-spatial" since users cannot arbitrarily locate themselves into the spatial environment—only along paths that the original video was recorded; hence, it is a limited spatial navigation interface. Tourgether360 gives users the ability to make sense of the video semantically—the minimap allows them to use architecture of the scene to make sense of, contextualize, and navigate the video data.

Figure 1 illustrates how two collaborators are embodied in the shared 360° scene together. Each collaborator's viewing position (time, space and orientation) is embodied by a viewing capsule, and they can position and mark points of interest for one another. This allows collaborators to experience the video as if they were embodied together in the video environment. While in the current implementation, mapping for Tourgether360 relied on manual parameter tuning, commercial applications are now able to reconstruct these scenes using 360° photographs[1], and computer vision approaches have shown the ability to reconstruct rich environments with only a limited amount of video data [39].

We evaluated our navigation approach with two studies. The first study involved 16 participants, and compared the embodiment and navigation elements of Tourgether360 versus a conventional interface for collaborative exploration of a 360°. We found that participants preferred the Tourgether360 interface due to its strong support for building mutual understanding. Based on the feedback from the first study, we developed the second version of Tourgether360, which provided participants with a mechanism to persistently gesture and refer to specific locations in the 360° scenes. We then conducted a second observational study with an additional 16 participants, where they collaboratively explored 360° using the Tourgether360 interface, completing tasks that required navigating through the video and identifying points of interest. We found that participants had very little difficulty adopting and using the pseudo-spatial navigation technique and used this to navigate the videos as opposed to using the timeline scrubber. Furthermore, we observed that participants were able to effectively use the cues to communicate and coordinate their interaction with the video.

We make three contributions in this work: first, we contribute extensions to a navigation technique for 360° tour videos that employs a pseudo-spatial approach to complement contemporary temporal navigation techniques; second, we contribute a system that realizes this technique and facilitates multi-user interaction; finally, based on our studies of Tourgether360, we uncover a number of new interaction challenges and opportunities based on pseudo-spatial navigation approaches that future designers ought to consider.

2 RELATED WORK

To set the stage for our work, we outline three related areas of research: first, we describe recent work that has explored new metaphors for navigating 360° videos, which consider spatial navigation approaches; second, we describe efforts to support collaborators exploring and using pre-recorded video streams, and then finally we outline some foundational work on collaborative virtual environments that inspired our approach.

2.1 Interaction with 360° Videos

Research exploring interaction with 360° has either focused on supporting orientation navigation (directing one's view in the video), or temporal navigation (controlling playback or directing one to interesting moments in the video). In terms of orientation navigation, considerable prior work has explored how to ensure the viewer does not miss important points of interest. Several approaches automate this through computational measures (e.g. [14]), while other researchers have designed mechanisms to signal to viewers where the view should be oriented. For creating pre-defined 360° stories, Pavel et al. [25] propose two techniques to orient a viewer's perspective when the playback comes to a pre-defined point of interest. On the other hand, Lin et al. [17] propose visualizations for points of interest that are out of the FOV of the viewer, allowing the viewer to see the point of interest in an inset video. Mäkelä et al. [18] show that such indicators of others' viewing interest (social indicators) can improve the experience, even if they are subtly distracting.

Several researchers have also proposed new techniques for temporal navigation of videos. For instance, Petry & Huber [27] explore multimodal gestures for playback controls for 360° within a head-mounted display viewing context. Similarly, Ruiz et al. [30] apply this approach within a multi-person viewing context. Neng and Chambell [21] present a 360° player that augments the traditional

[1] http://www.reconstructinc.com

timeline with cues representing points of interest, and regular thumbnails for some frames in the 360° video. For scrubbing through videos, VRemiere provides a "Little Planet" navigation technique for these videos, which can provide some spatial awareness [23].

Like our work, Route Tapestries is a departure from this prior work, focusing on blurring the boundary between navigating time and space [16]. In their work, they produce a navigation timeline that presents a slit-scan visualization immediately to the left and right of the forward-facing vector of the video. This timeline becomes a spatial "tapestry" of the scene, and presents the user with identifiable landmarks for navigating the video. This sort of "context" view is similar to the approach in Sight Surfers [24], which provides an overhead 2D map with trajectories that can be used for navigation. Our work builds on these approaches to enable a similar kind of spatial navigation that simultaneously manages the temporal navigation, motivated from the perspective of collaboration.

2.2 Collaboration Over Pre-Recorded Video

Studies of people navigating 360° together have revealed several communication problems that can be resolved through technology [18, 29, 34]. Particularly for experiences where viewers can watch simultaneously, there is a strong need for systems to provide an awareness of where others are viewing, and potentially gesture support to support communication and coordination. When collaborators can be at different points in the video at the same time, there is also a need to support this sort of temporal awareness. CollaVR provides this awareness, enhancing the timeline view with extra scrubbers, as well as through a colour-coded rectangle in the viewport that represents the collaborator's perspective [22]. Systems that support playback of VR recordings also provide this type of support, where the scene can be played back and viewed from different perspectives [36].

We were also inspired by prior work that has studied how people communicate about video recordings [4, 40]. Yarmand et al. present a study of YouTube comments, noting that while some comments make use of time-based references to the video, the majority of comments make reference to intervals in the videos—with specific reference to visual entities [40]. Aligned with this, Dodson et al. present a study of how lecture videos are explored within the context of recorded lectures, noting that content-based features (i.e., transcripts) were more useful for gross navigation compared to timeline navigation [3]. Thus, while time is an easy computation index into videos, people rely more on content-based features to communicate and navigate through videos.

Our approach centers on the insight that navigation through video may be better supported through a semantic, contextual understanding of the content (i.e., what is in the video) rather than time. In the context of 360°, we explored a visual map-based approach that provides this context view, combined with the normal view of the scene, which is the focus view.

2.3 Collaboration in Virtual Environments

Many of the problems outlined by researchers studying collaborative viewing of 360° videos [29, 33, 34] are reminiscent of early CSCW research focused on Collaborative Virtual Environments (e.g. [1, 2, 7, 9]). These virtual environments were designed to support multiple collaborators, and designers were forced to contend with basic awareness issues: (1) Where are my collaborators? (2) What can my collaborators see? (3) What are they looking at? (4) How can I draw someone's attention to what I am talking about? Early work by Benford identified many of these issues [1, 2, 9]. Follow-up work explored how virtual embodiments for collaborators could provide insight into presence and view orientation (e.g. [1]). Subsequent work identified ways in which viewports, and gestures could be supported, where some of these could exaggerate or create representations that might not mimic real life [5, 7, 8].

And, while some work has explored how to provide additional cues for awareness in 3D workspaces [32, 33], most embodiments we see in CVEs or video games have not pushed the boundaries of the embodiments first envisioned in the mid-1990s. Recent work has pointed to how some of these challenges with deictic references (and dereferencing) still persisting to this day [37, 38], with very few solutions addressing many of the challenges identified decades ago. Recent work has showed how these avatars and their presentation can be made to support effective mixed reality collaboration if some aspects of reality are ignored—e.g. size of collaborator [28].

We take inspiration from this prior work in laying down the user experience goals of our approach. While the domain of interest is slightly different (i.e., 360° vs CVEs), because we realize the 360° video in a CVE-like environment, many of the same embodiment approaches may still be applicable.

3 USER EXPERIENCE DESIGN

Our goal in designing Tourgether360 was to create an effective way to collaboratively watch 360° videos with others. These videos are characterized by a non-fixed position camera progressing along a path, and frequently focus on tours of tourist areas or cultural heritage locations. Building on prior work, we identified four major design goals:

- *DG1: Support semantic navigation of the video space.* Prior work on video navigation has demonstrated that temporal navigation can be meaningfully augmented with semantic navigation (i.e. with some understanding of the video content itself, e.g. [36]). Tourgether360 should allow people to navigate through the video data via semantically meaningful landmarks in the environment. This would obviate the need to think about or remember particular timestamps in the video. For instance, one can think, "I'd like to see the entrance of the cathedral," rather than needing to scrub through the timeline of the video to find the entrance of the video (in which case, one would need to control both the time scrubber as well as the orientation of the camera).
- *DG2: Support awareness of collaborators' perspectives and temporal position.* Watching 360° with others on independent displays means that collaborators may engage with the immersive experience independently, meaning that collaborators will separate—both in terms of their viewing perspective, as well as in their temporal navigation of the video [33]. Knowing what others are looking at, and where they are is an important part of feeling co-present [10, 11].
- *DG3: Enable smooth engagement and disengagement with collaborators' perspectives.* We know collaborators sometimes like to explore the video independently, in a loosely-coupled mode of interaction. At the same time, having a shared perspective supports smooth, detailed coordination and conversation in a tightly-coupled mode of interaction. A tool should allow collaborators to smoothly move between these modes of interaction [33, 34].
- *DG4: Support deictic reference with a semantic understanding of the environment.* Finally, the tool should allow collaborators to point and refer to things in the video and environment; further, these should be somewhat persistent given that collaborators may not be looking at the same thing all the time [37]. In practice, we know that many deictic references only make sense in the context of a conversation [11]. We have also seen this in some of the earlier groupware systems (e.g. [11]), where designers created "telepointers" to allow people to make deictic references. In those cases, we would say that the researchers chose to create non-persistent deictic references to support their synchronous interactions. In Tourgether360, our thinking was that users would find themselves frequently viewing the different part of the environments; as such, we imagined they would want to leave "landmarks" in the environment for others to find. Thus, our choice was to create more persistent references–though these could be removed.

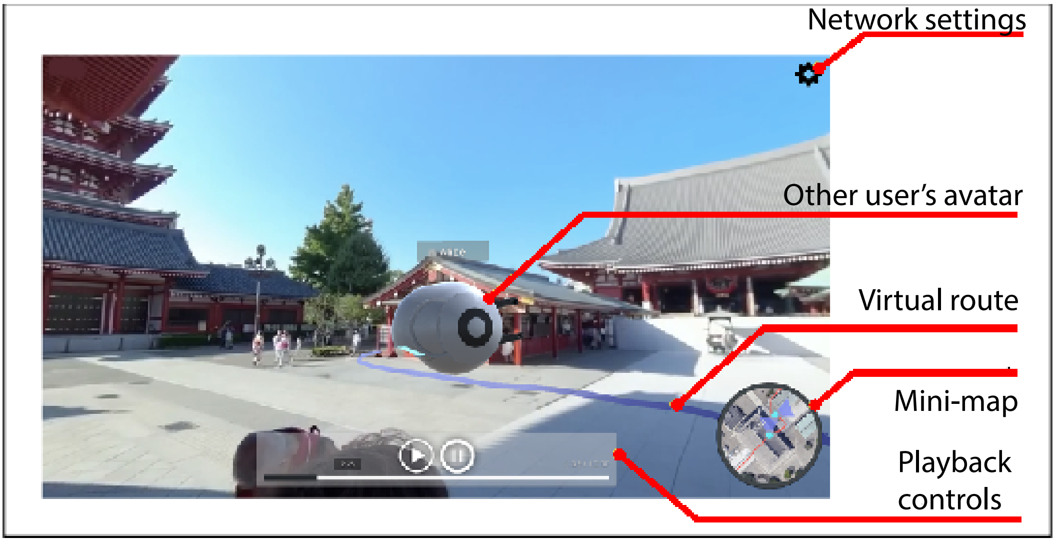

Fig. 2. Full Interface of Tourgether360 from the perspective of a user (Bob).

We executed on our approach by reconceptualizing 360° as a shared virtual 3-dimensional space and implementing a number of unique features that increase users spatial understanding of the environment and enable allocentric navigation. These features include pseudo-spatial navigation mechanisms, user embodiment and pseudo-spatial annotation, and allocentric coordination affordances. We realized our design vision in the experimental video player called Tourgether360. The interface of the software is presented in Figure 2.

In the following paragraphs we outline how each major function of Tourgether360 addresses the design goals above.

Pseudo-spatial navigation. Inspired by prior work [24] and the UI of the 3D video games, navigation affordances in our application are supported by an overhead schematic interactive minimap of the 360° video tour environment (Figure 3). The minimap provides a virtual path that represents the route on which the tour takes place, where the user's position in the video is represented by a blue dot in space. Depending on where users are currently positioned, their avatars are mapped on the corresponding place on top of the virtual path. The users' gaze direction is also represented on the minimap by the conical light beam from the avatars. The minimap allows users to navigate through the video using spatial landmarks visible from the minimap (DG1).

The minimap serves the users as a navigation control mechanism. By clicking and dragging the mouse along the virtual path on the minimap, the users effectively scrub the video timeline back and forth, rewinding and forwarding the video. In addition, the users can scrub the video by scrolling the mouse wheel.

As illustrated in Figure 4, the metaphor of the 3D space is also realized in the main video view. Here, the path of the video is represented by a continuous line stuck to the ground, showing the forward and backwards route of the tour from the first-person perspective of the user. This virtual route served as a visual representation of the actual route in the environment through which the person with the camera moved while the video was originally recorded. Thus, virtual track provided users with an understanding of what physical places the users pass when playing the

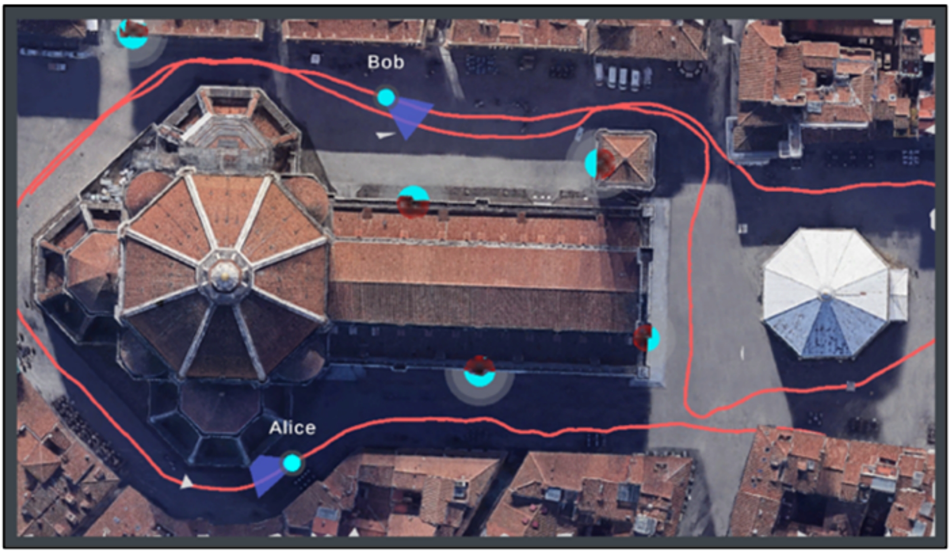

Fig. 3. Minimap of the Florence Cathedral environment shown in one of the 360 videos we used in the study. Here, the path taken in the 360 video is represented by the red line. Alice and Bob are at different parts of the video, where their viewing orientations are represented by a cone. Finally, spheres represent marked "points of interest" that were placed by the collaborators.

video. To enhance the illusion of being present in an environment with geometry, parts of the path are cropped if they could not be seen around the geometry of the space.

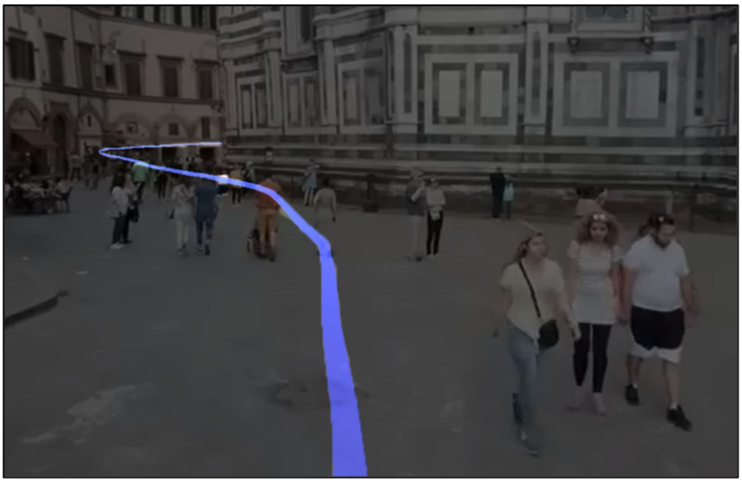

Fig. 4. Representation of a virtual route overlayed on top of the video. Taken from Alice's point of view (from Figure 3), the path (highlighted blue) illustrates how the video tour will take them around the cathedral. Because Tourgether360 understands the architecture of the space represented in the video, the line path is cropped at the edge of the cathedral.

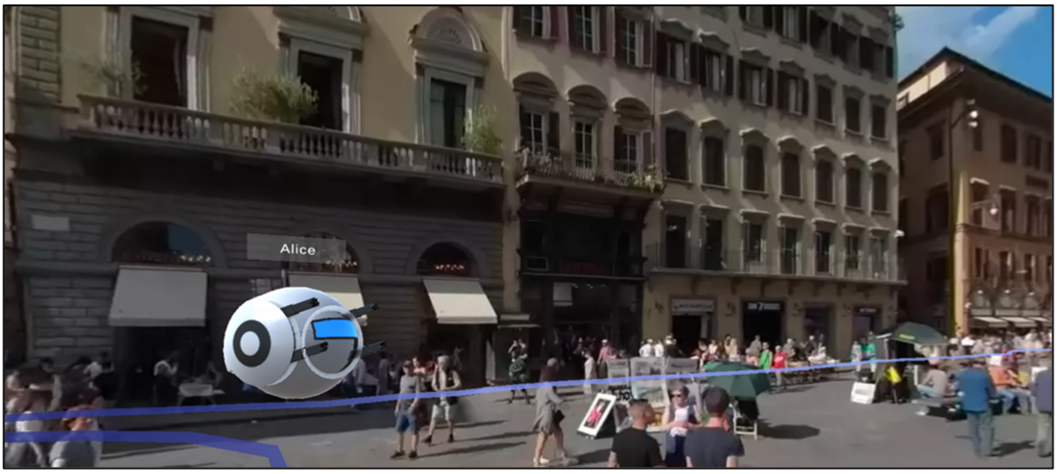

Fig. 5. Representation of the user avatar overlaid on the top of the played video on the virtual route line.

Collaborator Embodiment. As illustrated in Figure 5, each collaborator is embodied by an avatar in the 3D video tour scene. The avatar is a flying spherical robot with four antennas indicating the "face" part of the robot. This "face" is synchronized with the user's camera's forward vector to indicate the gaze orientation. This embodiment approach in the 3D scene provides awareness of others' temporal position and view orientation (DG2). The embodied avatar is shown in Figure 5, and reflects a whimsical narrative, where users are using spherical robots to "time travel back to this location." Although these robots are visually sterile, the design choice also reflects the data characteristics we have for each user—that of their temporal/spatial location in the video stream, as well as that of their viewing orientation. We specifically avoided anthropomorphic embodiments, as limbs might have implied that users would be able to gesture with them, or to influence the environment.

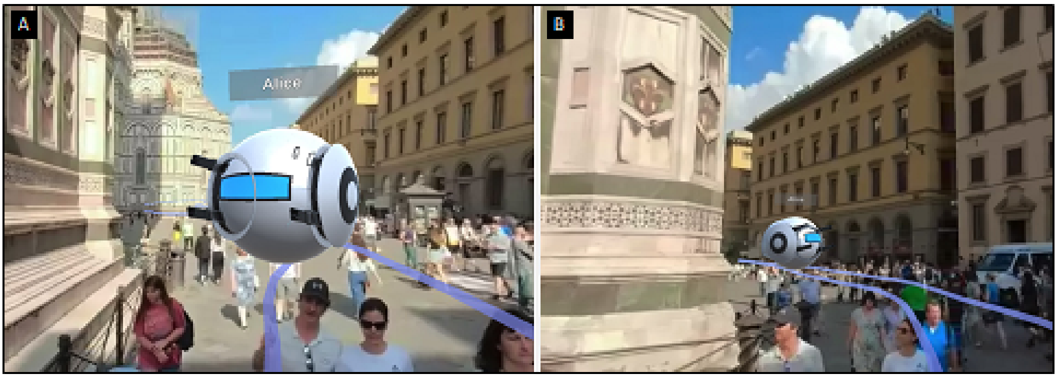

Fig. 6. One user (Bob) sees his partner's (Alice) avatar while playing the video. (a) Both users are close to each other in the video (b) Alice pauses the video, while Bob continues to play it, thus seeing Alice's avatar gradually becoming farther and smaller as she stays in place.

When collaborators are watching the video together, the apparent distance between two avatars in the 360° scene is equivalent to the temporal distance between two collaborators. Thus, as

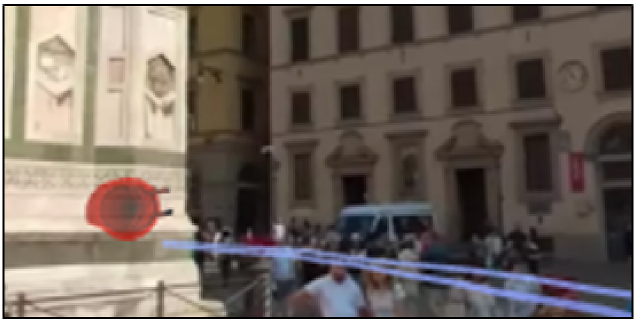

Fig. 7. A collaborator's avatar is rendered as a silhouette if they would normally be occluded by the environment (here, by the wall of the building).

illustrated in Figure 6, if one user pauses the video while the other continues to watch, the former will see the latter's avatar gradually going away and shrinking in size as it moves farther down the route in the video tour. As illustrated in Figure 7, to maintain the illusion that collaborators are navigating an environment rather than a video (DG1), collaborators' avatars are presented using a silhouette representation if they would normally be occluded by the architecture of the space.

Collaborators can use the embodiments to engaging and disengaging smoothly with each other through view synchronization. A user can also assume a spectator role by double clicking on another collaborator's avatar, which synchronizes both collaborators so that both playback and view orientation are synchronized to the guide. When the leader now navigates through the video or changes their orientation, the other user's view also changes. Users can regain control by simply moving their orientation or explicitly navigating once again. This allows them to move in and out of engagement with one another smoothly (DG3).

Pseudo-spatial annotation and allocentric navigation. As illustrated in Figure 8, users can annotate the environment and navigate around it using external artificially created landmarks – pseudo-spatial markers that are placed by the users directly in the environment of 360° video tour during watching. We introduced this feature after the first round of iteration, following the

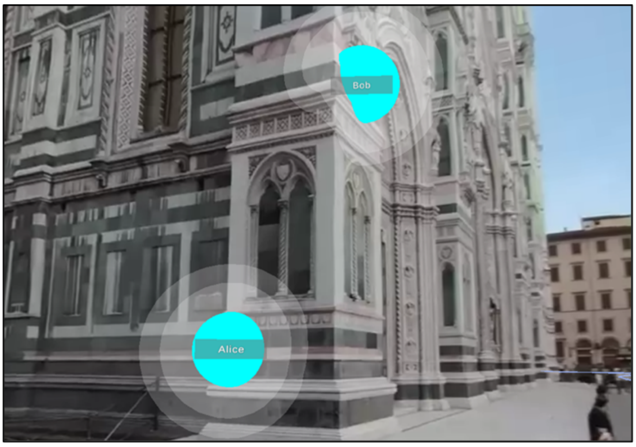

Fig. 8. Representation of pseudo-spatial markers placed on the walls of the Florence Cathedral by two users (Alice and Bob).

results of Study 1 (Section 5). The markers are visible to everyone and, from the user perspective, diminish or increase in size depending on the closeness of the user to them. This allows users to communicate about the environment via deictic reference (DG4) and reinforces the notion that the annotations are about the semantic space (DG1).

Users can instantiate markers by double clicking at the point of interest in the environment where they want to place them. Clicking on a marker will teleport the users to the point of interest and time in the video when this marker was instantiated. By default, the teleportation brings a user to the timestamp when the marker was created, and immediately reorients the user so the marker is centered in the view. Users also can delete the existing markers by double clicking on them. We added this mechanism partly based on the feedback we received from Study 1.

4 SYSTEM ARCHITECTURE

We created Tourgether360 using the Unity game engine environment using the Multiplayer Networking library (MLAPI). In the following paragraphs, we describe the implementation of system's main technical features.

Route Extraction. The virtual route was extracted using Simultaneous Localization and Mapping technique (SLAM), specifically through an open-source implementation of ORB-SLAM [19] using the omnidirectional camera model. The SLAM algorithm takes a monocular 360° video as its input, detects a set of static geometric feature points from the environment shown in the video, and computes the camera position with regards to the detected feature points in each video frame. Camera paths returned from SLAM typically involve some high-frequency jitters. As we did not need highly precise paths, we simply sampled the returned positions at a 0.5s interval and connected them as the camera trajectories.

Visual SLAM algorithms are reasonably robust to dynamic objects (e.g., pedestrians and vehicles), but they may fail when the environments are not well-lit or lack enough distinct features (e.g., a hallway with walls that are purely white or covered with repetitive patterns). Many virtual tour videos show visually complex scenes such as city landscapes and thus contain sufficient features for reliable tracking.

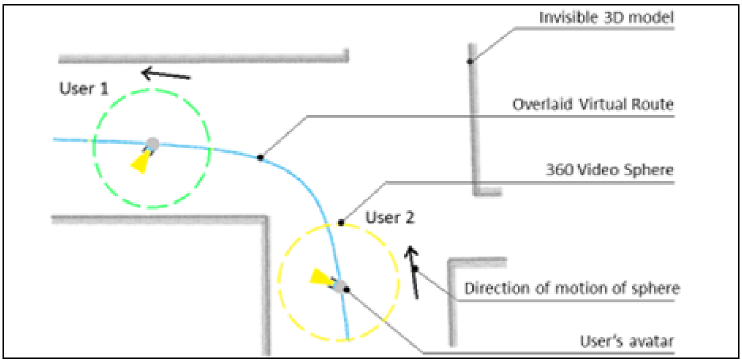

Fig. 9. Overhead diagram of two users (and their video spheres) moving along the virtual route in the environment. Here for user 1 only the green sphere and the avatar of user 2 is rendered and vice versa. Note: The spheres are made small for visualization purposes, in the actual system they engulf all of the 3D model and the route.

Route Alignment. The SLAM process only detects sparse feature points of the environments. To achieve correct occlusion effect, we manually align the extracted camera path with triangle

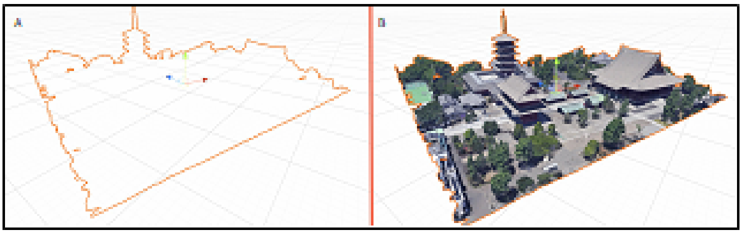

Fig. 10. 3D model of Asakusa Shrine Complex that we used as a virtual overlay for the corresponding video used in the study. (a) The model with a custom transparent shader that is used as a direct video overlay (traced for clarity), (b) The textured model used as an overlay for the minimap.

mesh models of the environment. Note this model is only for occlusion detection and thus not rendered (Figure 10). We used RenderDoc[2] to extract Google Map's 3D buffer cache of the location and converted the data to 3D models using Blender[3]. We then scaled, rotated, and translated the model to align it with the camera path in Unity, using the spare feature points obtained from SLAM as references. For each video, extracting the 3D model from Google Earth and aligning it took about 20 minutes. Future work can explore automatically downloading models based on geographic locations of the videos and aligning them using point-set registration techniques such as Iterative Closest Points.

The 360° video is rendered on a sphere around each user, which engulfs the 3D models, the virtual route, and the avatars of other users. On each client, only the sphere corresponding to the local player is rendered. Each sphere along with the parented user's avatar at their center, move independently along the fixed route in the unity space (Figure 9). This combination of the video and 3D models provides dynamic occlusion of users' avatars, and the virtual route. For instance, if the other user's avatar moves behind a wall, the avatar changes to a silhouette-like appearance. The implementation of this function was via a custom shader, which although is transparent (to allow for the unobstructed view of the video), tints the shaders of specific objects like avatars to a red fresnel (silhouette) shader when occluded (e.g., Figure 7). Similarly, the path is occluded when it is obstructed by any solid spatial entity (e.g., a wall or a building in the video, as in Figure 4).

Pseudo-Spatial Markers. The pseudo-spatial positioning of the markers in space was implemented via a ray-casting technique, where a ray from the mouse cursor points on the screen determined the position of the marker in the location where this ray hit the 3D model of the environment. When markers are created, the markers are instantiated and positioned directly on the 3D model. Users perceive the elements as if they are synchronized with the actual video, appearing to stick to the place where they were instantiated. This is illustrated in Figure 7 where markers appear to be stuck to the actual walls of the building and get large/small depending on the user's temporal and spatial distance to them. Markers which should not be in direct view from the camera (example, being obstructed by a certain face of a building) are occluded by the hidden 3D model.

Use of Tour Videos in Outdoor Environments. In our testing, we found that tour videos of outdoor environments worked best for our approach. First, outdoor videos tend to be better lit than indoor videos; the occasional poor lighting in indoor videos would cause problems for our SLAM implementation, resulting in sometimes bizarre movement paths. Second, because the distance from the camera to the visible features of the environments (i.e. walls of buildings) tends

[2] https://renderdoc.org/

[3] https://www.blender.org/

to be far compared to in indoor environments (i.e. objects in the environment, or even indoor walls), features are generally visible in more frames and thus enable more reliable tracking . Finally, outdoor environment geometries were more readily available from public sources compared to indoor environments. Given these constraints, our current implementation runs primarily on outdoor environments. In principle, advancements in computer vision technologies combined with an on-board depth camera for indoor environments could produce 360° videos that would work for Tourgether360, though our primary aim in this work was to explore the utility of the interaction and navigation techniques as described in our design goals.

5 STUDY 1: UTILITY AND USABILITY OF MINIMAP AND EMBODIMENT

To assess the utility and usability of the minimap and embodiment approach of Tourgether360, we designed a comparative lab evaluation where pairs of participants alternately used Tourgether360 and a control interface to discuss locations in a 360° tour videos, and solved basic problems in these videos. The purpose was to understand whether the minimap and embodiment approaches of the first version of Tourgether360 (which did not have the annotation/pointing features) would be useful for collaborative viewing compared to a conventional 360° video experience. While connected via a video conferencing tool so they had full audio, participants each viewed the 360° videos from their own desktop or laptop computers, connected using the Tourgether360 tool (or the control interface). We asked participants to explore the videos while completing collaborative tasks together.

5.1 Study Design

Pairs of participants completed tasks using both the Tourgether360 interface and the control interface in a within-subjects design. We alternated the presentation order of the interfaces across pairs in a simple Latin square design.

5.2 Materials and Tasks

Interfaces. In the Tourgether360 condition (*Tourgether360*), participants used the interface illustrated in Figures 1 and 2, complete with the minimap and avatar embodiments. This version of the interface did not yet have the annotation/gesture features. In the control condition (*Baseline*), participants used a conventional 360° viewing interface, where they could scrub the timeline to move through time, and then grab the main view to swivel the camera around. Their partners were not visible as avatars, and there was no minimap. To make the two conditions slightly more comparable, we visualized their partner's position in the timline via an additional "scrubber dot." This gave participants an understanding of where in the video timeline the other participant was. This "scrubber dot" is illustrated in Figure 11.

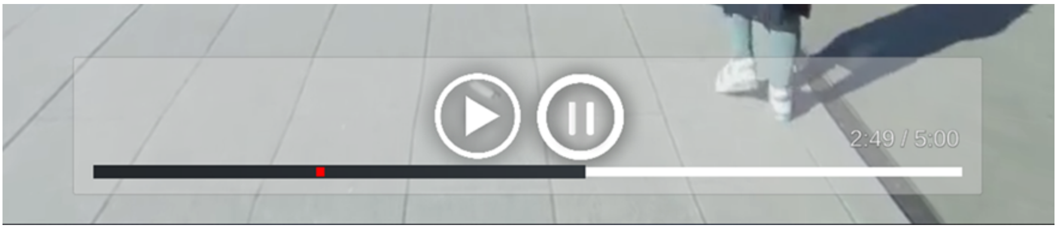

Fig. 11. The baseline interface timeline, where the other participant's position is shown as a red "scrubber dot".

Videos. We selected two publicly available 360° videos. Both videos involved a person holding an overhead 360°-camera capturing the environment, moving along a path. In each video, the person holding the camera was not visible, creating an illusion of a first-person view of the environment for users. We chose the videos for novelty (we did not want participants to be familiar with the locations), content distinctiveness (we wanted the content to be interesting and contain enough interesting objects for the tasks), and technical feasibility (3D models are available online). Each video represented a virtual tour of a specific popular tourist location: Florence Cathedral (duration: 15:10), and the Asakusa Shrine Complex (duration: 10:00). Both videos include static architecture and a large number of dynamic objects, such as pedestrians and vehicles. For our user study, we used the first ten minutes of the Florence Cathedral video, and the ten-minute Asakusa Shrine Complex video. Each of these two were split into two five-minute segments (one five-minute segment for task 1, and the second five-minute segment for task 2). Participants always saw the Florence Cathedral video first, and the Asakusa Shrine Complex second. We used different videos for different tasks to keep participants engaged and prevent them from memorizing the content. We counter-balanced the presentation order of the interfaces between participants.

Tasks. In this study, participants were to imagine they were viewing potential vacation destinations and planning out their trip. The study comprised two major tasks to mimic different kinds of collaborative activities with such videos: identifying then discussing locations, and extended study of the video locations. In Task 1 (identify and discuss), participants were given five minutes to individually identify at least two locations in the video they thought would be great places to take a photo together as a keepsake, and then were given an additional five minutes to share and discuss these locations to identify the single spot they would take the photo together. In Task 2 (extended video exploration), participants were given two tasks that required them to study the duration of the video (e.g. “Count the number of bridges.” or “Identify the most popular store.”). Participants completed both Task 1 and Task 2 with a single interface on a given video before redoing the tasks with the second interface on the second video. The tasks we chose focused primarily on locations that the video passed by, rather than on impromptu instances of dynamic activity (e.g. a passing truck).

5.3 Procedure

The study was conducted in three stages: introductory, training, and the study. At the introductory stage, we connected with both participants remotely using Skype or Zoom video conferencing software. The participants were shared links to the written description of the study and the pre-study demographic questionnaire. Once we had verbally reviewed all aspects of the study, including our data collection and handling procedures, we obtained verbal consent to participate. We then invited participants to run the previously downloaded application and connect to the dedicated server.

Once participants had joined the shared server, we asked them to start with the training video for their interface. For instance, with the Tourgether360 introductory video, researchers explained each aspect of the interface and all possible actions and operations possible while running the video, including operating the minimap, and so on. Participants were then invited to engage in the practice session and play with the app together until they felt completely comfortable with using the software. Once the participants had mastered all the required interactions, they were invited to complete the two tasks with the first interface. Participants then completed a questionnaire reporting on their experience. Then, they repeated the procedure with the second interface. At the conclusion of the second questionnaire, the researchers conducted a semi-structured interview with the participants, focusing on the various aspects of communication, coordination, and overall user experience of the interaction. This study protocol was approved by our Institutional Research Ethics Board.

5.4 Participants

We recruited 16 participants in pairs for the user study. Participants' age was between 18 and 35 years old, four identified as women, twelve as men. Thirteen participants were frequent video game players, with the majority of them playing at least weekly, although only eight of them have experience with video games involving 3D environment navigation, such as Call of Duty, and Counter Strike titles. In addition, 12 out of the 16 participants indicated familiarity with 360° videos, but most of them were not heavy consumers.

5.5 Data Collection and Analysis

We collected the field notes and logs of the participants' interactions with the software. Prior to the study, we also asked one of the participants to share their screen with us via Skype or Zoom, which we recorded. In the questionnaires, we collected System Usability Scores (SUS), a Shared Cognition assessment [36], and the Shared Presence data [15]. The questionnaire design was adapted from prior studies [15, 36] on sharing experiences in social VR. Shared Cognition measures how well participants understand each other within each pair. Shared Presence gauges participants' sense of being in the same space with the partner. The questionnaire can be found in our supplementary materials. In addition, we collected post-study interview data. Our analysis was grounded in our observations of participant behavior during the study and guided by their interview responses. We took a thematic analysis approach, grouping this data based on thematic relatedness, drawing common stories about participants' experiences from the data.

5.6 Results

We present our analysis results of post-study questionnaire and interview data. Overall, results showed that Tourgether360 was easy to use and preferred by participants because of its support for mutual understanding. We also identified a need for persistent spatial marking.

5.6.1 *Results of SUS, Shared Cognition, and Shared Presence.*

We calculated the SUS score and the mean response values of Shared Cognition and Shared Presence for each participant. We used similar statistical methods for the three metrics. Specifically, to account for the interdependence of responses from members in dyads, we modelled participants' responses using multi-level linear

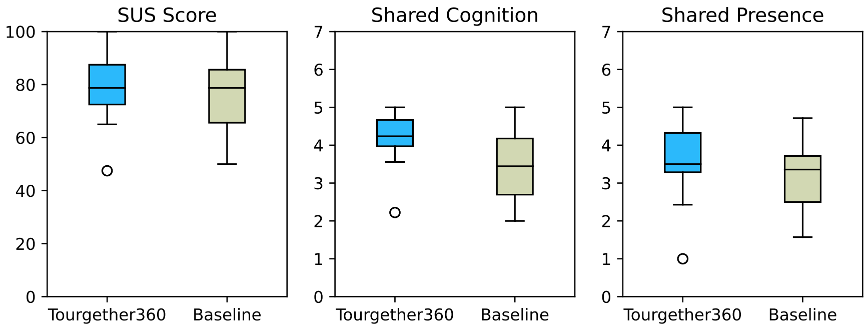

Fig. 12. Post-study questionnaire responses for Tourgether360 and the baseline interface in box-plots. The circles show outliers.

models, treating each pair as a cluster and the condition (Tourgether360 vs Baseline) as a fixed factor. The analysis was conducted using the package `statsmodels` in Python.

the SUS scores that participants gave for Tourgether360 (M=78.4, SD=13.2) were similar to the scores for Baseline (M=75.3, SD=15.2), suggesting that they found Tourgether360 as easy to use as traditional video player interfaces (Figure 12).

As shown in Figure 12, participants reported significantly higher levels of Shared Cognition ($p<0.01$) with Tourgether360 (M=4.21, SD=0.71) than with Baseline (M=3.45, SD=0.94). Further analysis on individual questions showed that participants found it significantly easier to understand their partners (Q2.1), refer to objects (Q2.5), and locate objects mentioned by others (Q2.6). Our analysis did not find a significant difference in reported Shared Presence between Tourgether360 (M=3.63, SD=1.00) and Baseline (M=3.21, SD=1.00). We speculate that this may be because these items are drawn from prior work that considers the impact of the fidelity of others' representations in the space (e.g. fidelity of an avatar in VR applications), and so this measure was not sensitive enough for our purposes.

5.6.2 Finding 1: Participants preferred Tourgether360. When discussing their experiences of performing study tasks, participants overall commented more favorably on Tourgether360 than Baseline. Echoing with the Shared Cognition data, participants highlighted that the minimap and user embodiments of Tourgether360 made communication and coordination smoother. P11 described how Tourgether features reduced the effort for establishing common ground with his partner, *"Looking at the minimap I can make sure we are looking at the same thing, but with the other (video) player I constantly need to refer to specific objects to confirm our views are in sync"*. We observed this pattern on other pairs as well, who frequently asked each other questions such as *"Do you see a small temple?"* [P14] or *"Are you looking at [restaurant name]?"* [P9] while using Baseline. Further, participants found it helpful that when common ground is in question, where Tourgether360 allowed for synchronizing their views immediately. They also liked the ability to guide their partners directly with allocentric references, such as "look to your right", with Tourgether360. While the Shared Presence metric did not show a significant difference, several participants mentioned that seeing the robot embodiments enhanced their sense of being 'in the same video' with their partners, as P3 described *"It felt like, I'm sitting in a room with my partner and we both have become a part of the video."*

5.6.3 Finding 2: Participants need a way to refer to locations in the scene. When asked about improvements to Tourgether360, several participants suggested adding a tool for them to precisely point to objects in videos. We noticed that while it was relatively easy to synchronize views in videos using Tourgether360, referring to individual objects still required lengthy, explicit verbal descriptions. Meanwhile, participants expressed their desires to bookmark specific locations for later references. P14 explained her rationale, *"If I want to record a location so that later I can study it again or point it to my partner, I have to note the timestamp down. It would be nice if I can just bookmark locations."* These observations and suggestions prompted us to devise the pseudo-spatial marker, which can serve as both deictic indicators as well as persistent annotations.

Based on this finding, we developed the fourth design goal (DG4: Support deictic reference with a semantic understanding of the environment), and added mechanisms for participants to persistently refer to objects and locations in the video.

6 STUDY 2: TOURGETHER360 FOR COLLABORATIVE EXPLORATION

Based on the findings from Study 1, we revised and improved the interface of Tourgether360 to include an annotation and gesturing mechanism (described above). Study 1 convincingly demonstrated to us that the navigation approach in Tourgether360 was effective and desirable to participants compared to the Baseline. To this end, we were now more interested in a more holistic perspective on the experience when pairs work together with a more extended, open-ended task. That is, how would people use the interface to view videos together? Would they work together and stay together, or would they work independently and try to maintain some awareness of one another during the collaborative tasks? The second study was focused on understanding the overall experience of the interaction, to uncover behavioral and collaborative patterns, and to understand what functions participants find particularly useful during the study.

6.1 Study Design

In this study, the participants only had one condition—the full Tourgether360 interface.

6.2 Materials and Tasks

Videos. In this study, we used the full Florence Cathedral and Asakusa Shrine Complex videos, and added two additional videos: Rome's Colosseum (duration: 4:07), and the Asakusa Market area (duration: 6:00).

Tasks. The study tasks were designed to encourage the participants to work in both tightly coupled and loosely coupled modes of interaction. In each of the tasks, the participants were free to use any functionality available to them in the software in whatever ways they deemed appropriate, communicate freely, and work in whatever style they wanted. The tasks were limited in time: the duration for each task was 60% of the duration of the respective video used for the task. This was done to encourage some loosely coupled work, where participants could work independently of each other to cover more area of the location quicker, in addition to the tight coupling collaborative mode.

In the first task involving the Asakusa Shrine Complex, the participants were asked to locate the biggest shrine and find several points where they thought that the view on the shrine was the best. In the second task involving the video of the Colosseum, the participants were asked to identify several points where they would like to take a picture together. In the third task with the Florence Cathedral, the participants were asked to find the place with the best perspective on the Cathedral and its entrance.

6.3 Procedure

We followed an identical procedure as in Study 1, except there was only the Tourgether360 interface to introduce. The training video was amended to include all possible actions including the creation and deleting of markers. This study protocol was approved by our Institutional Research Ethics Board.

6.4 Participants

We recruited 16 participants in pairs for the user study. Participants' age was between 18 and 25 years old, four identified as women, twelve as men. Eight of the participants were frequent video game players, with majority of the users playing at least weekly, and having experience with video games that involve navigating in 3D environments. In addition, all participants indicated familiarity with 360° videos, but stated that they rarely engaged in watching such videos. No participants had prior experience with Tourgether360, and none were involved in Study 1.

6.5 Data Collection and Analysis

We collected identical metrics as in Study 1, though this time in our analysis, we focused on the data from our field notes as well as the recorded log data from the studies.

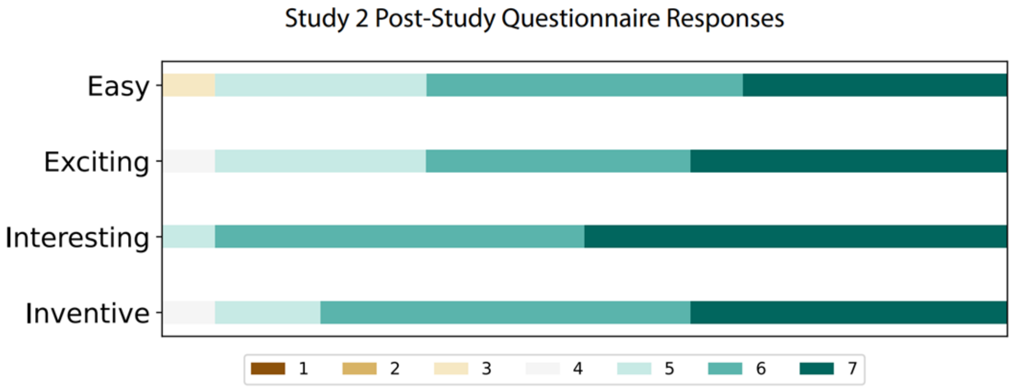

Fig. 13. Study 2 post-study questionnaire responses. Questions asked included 'I found the experience easy/exciting/interesting/inventive'. 1-7 indicates 'strongly disagree' to 'strongly agree'.

6.6 Findings

Participants reported having very positive experiences with Tourgether360 during the study. Without exception, participants were able to learn and use all the functions of Tourgether360. All the users quickly and confidently navigated the video tours, creating collaborative markers, and using the video timeline slider and the minimap for coordination and communication. In the post-study user-experience survey (Figure 13), the participants rated their overall experience as quite easy (5.9 on the scale of 7), interesting (6.5 on the scale of 7), exciting (6 on the scale of 7), and inventive (6.1 on the scale of 7). The most salient aspects of participants' behaviors emerged around participants' perceptions of the social aspects of the experience, their use of spatial navigation using Tourgether360, and their use of the pseudo-spatial markers for communication and coordination.

6.6.1 Collaborative Coupling in Task Completion. The pairs in the study demonstrated a wide variety of collaborative coupling styles in their work. In general, it was challenging to classify pairs as working in strictly one style or another; rather, their work typically involved a wide range of coupling styles—even within the execution of a task with a single video. Figure 14 illustrates a subset of charts that map video playback time as a function of task time. Normally, the trajectory of each plot moves diagonally upward (i.e. normal passage of time). When participants have paused the video, that segment of their trajectory in the chart is horizontal (task time continues, while video playback pauses); large jumps represent moments when a user has clicked on a landmark, or has decided to follow their collaborator, and finally, jagged edges represent when a user was "scrubbing" through the video (either on the timeline, or on the spatial path in the focus view).

These charts are helpful to understand how groups worked together or independently. Figure 14 (A) shows Group 5's explorations in video 4 (Florence). Here, we can see that [P9] and [P10] stay very close to each other, essentially viewing the video together through the entirety of the trial. This contrasts with Figure 14 (B), where Group 4 (also in video 4) spend their time quite differently. [P7] goes ahead in the video, and lays down several landmarks, and then works backwards, eventually ending up at the beginning of the video. At this point, [P7] and [P8] go through to the end of the video together, discussing the landmarks together, and bouncing back and forth using the landmarks. Figure 14 (C) more clearly show how Group 6 use the landmarks to navigate video 3 (Shrine). Here, [P11] jumps ahead and begins at about halfway through the video, and lays markers down; once he is done, they both begin using the landmarks to navigate to different interesting

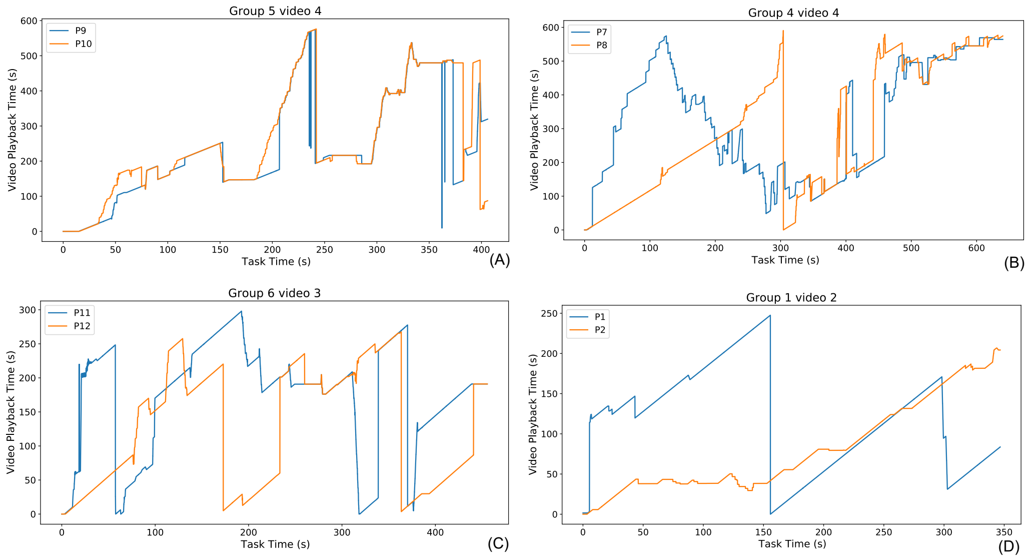

Fig. 14. Mapping between task time and video playback time in four example study sessions.

points in the video. For instance, there seems to be two markers at 180s and 225s in the video that both participants return to multiple times. Finally, Figure 14 (D) shows Group 1 in their exploration of video 1 (Colosseum). Here, they are working in a very loosely coupled fashion—only spending about half the time watching "together". [P1] goes ahead, halfway through the video, and begins his work there, whereas [P2] spends the first part of the video exploring something interesting that happens around the 45s mark. [P2] does end up putting down a marker that [P1] uses to navigate to see the event (which he does at 300s into the task time).

The groups did not seem to stick with one coupling style or strategy across the videos; instead, for the most part, they blended coupling styles to suit the task or their preferences in the moment.

6.6.2 Perception of the Experience. Participants' overall descriptions centered on the highly social experience provided by Tourgether360, and how it resonated and reflected their experience with video games and virtual world. We refer here to participants by participant number in Study 2 only.

Tourgether360 as a Social Experience. Participants described the experience fundamentally social, and engaged in the study tasks as if they were in a shared virtual space. For instance [P2] reported, *"We could see what other people are doing, so it was sort of a social experience and not just a video"*. The embodiment of participants' partners within Tourgether360 played a significant role in participants' perception of the social character of the experience: because the participants saw each other as the avatars the scene, they referred to objects in the scene in relation to each other's avatars. For example, in one instance, [P3] suggested to his partner: *"[P4], the building to your left could also be the main temple"*. Groups relied on the embodiments to tell them whether their partners were "with them" or "nearby", and adjusted speech and conversation accordingly, where they would beckon one another in different ways (e.g., *"Come over here"* [P4] when distant, versus, *"Look at this"* when nearby). One pair [P9 and P10] took this to the extreme, where they ensured that they stayed in each other's view throughout the entire study. This would ensure they could see each other, and what each was looking at as they "walked" through the space in the video—a sort of visual confirmation that they were "with" each other as they went through the video.

Participants also enjoyed the ability to "synchronize" views with one another like a guide-follower pattern. When a participant would click onto their partner's view (i.e., to become a follower), the partner would then temporarily become a tour guide, showing specific points of interest for discussion. We observed every group do this at least once; furthermore, many groups would fluidly flip-flop these roles (as well as disengage) during the study session depending on the situation. For example, one participant [P10] requested his partner [P9] to sync his view to show several relevant locations. In the process, [P9] mentioned another angle which [P10] had missed. Immediately, [P10] synced to [P9]'s view to see this location. Similarly, P3 had his partner synchronize his view before showing each point of interesting to assesses its relevance [P3] before synchronizing his view to his partner's one to watch see the locations that his partner had found.

Tourgether360 as a Video Game. Participants reported that Tourgether360 felt more like an immersive video game than a 360 video player. When asked to compare their experience of Tourgether360 with other computer-related activities, [P6] reported, *"This reminds me mostly of video games. Although we watch video, the interactivity of the process makes it not like any other video viewing app, and I feel that we are playing rather than watching it."* This suggests that the navigation mechanism, albeit constrained by the nature of the video (i.e., along a track), allowed sufficient latitude and flexibility to give participants the feeling that they were moving through an active environment. Similarly, [P1] reported, *"It was sort of an interactive video and not just a regular video"*, which is consistent with [P9] who described it as, *"It was a mixture of both, a video watching experience as well as playing games, because although I was watching a normal video, simultaneously I was playing around with the app as well."* Other participants reported that the experience was akin to online multiplayer games with virtual environments (e.g., first person shooter games)—even if they had relied primarily on the timeline slider to navigate time. As explained by [P11], the very presence of others (through the embodiments) helped to create this impression and perception, *"[The experience] felt similar to playing Role Playing Games (RPGs), since I see other users directly, and we can move close to each other or move away"*.

The participants' understanding of the experience as an interactive video game was reflected in how they navigated the videos. All groups except one looked at the videos non-linearly, jumping back and forth to different places in the space (and video) to explore the environment. For example, one pair used the minimap at the start of the study session together and saw that the most interesting parts of the environment were located toward the end of the video [P1, P2]. One of the participants jumped straight to this final piece of video and then spent most of the time of the study there [P1]. The participants also engaged into playful interactions using the interactive functionality available to them. Several participants at some point started to jokingly delete each other's markers and laughing about it (e.g., [P3, P4]).

Another common framing of the experience was participants' comparison of the interaction to being present in the real world. [P8] reported, *"I don't even need to go outside. I can go and browse the world with my friends however I want without even coming out of my desk"*. Participants were excited by the potentially being able to use the app when planning to go to the real locations in the future. For example, [P10] suggested that a potential use case would be to go to the location in Tourgether360 and mark all of the interesting parts so that he would know where to look ifhe was to actually travels there.

6.6.3 Preference for Spatial Navigation and Wayfinding Behaviors. Participants were able to quickly appropriate the spatial navigation metaphors presented by Tourgether360, and we noted that participants generally used spatial navigation strategies rather than a temporal navigation approach.

Tourgether360 gives users two primary ways to navigate and wayfind through the video: the minimap, and the timeline slider. The minimap gave participants the ability to navigate through the video relying on a spatial mental model (e.g., "near the main building in the Shrine Complex", or "at the front of the building" and so on); on the other hand, the timeline slider represents a traditional, time-indexed navigational model of the video (e.g. "at the start", "two minutes into the video", and so on).

While we observed participants using both tools to navigate through the video, participants overwhelmingly relied on spatial navigation strategies—both in terms of which UI elements they used to navigate through the video, and how they communicated with one another. All but three used minimap as the central hub of their navigation and wayfinding activities, rarely ever using the timeline scrubber. They even referred to architectural elements in the minimap and relying on spatial relationships when communicating with one another (e.g., [P15] *"Go to toward the front of this temple, there is the good spot"*; [P10]: *"We'll get a better view from the front side of the church."*; [P11] *"At least one location is fixed for the photograph, the one from the front side of the shrine."*). Similarly, when [P1] was commenting on his partner's work, *"The markers that you have placed from the side of the building are also very good. Photographs from this angle would be great"* [P1]. Yet another user explained the particular time within the video by using spatial reference instead of time: *"The place where it goes in [the alley], right?"* [P12].

The minimap was also a way that they understood each other through the avatars, the route of the tour represented by the path, and the markers that they put into the environment. In many cases, participants reported mainly looking at the minimap to make sense of their ongoing activity, rather than the video. For example, P2 reported, *"[I] mostly looked at [my partner] on the minimap and not directly in the video because it was easier to understand precisely where he is now"*. The minimap allowed a type of precision in how they navigated to locations in space: [P7] reported enjoying the scrolling ability as it allowed him to reach precise locations of interest, since he could understand in detail where the location was exactly and jump to the corresponding route point in the minimap. Only three of sixteen participants predominantly used a temporal navigation strategy [P3, P8, and P9]. They referred to their own locations and locations of the markers that they put into the environment as the time stamps in the video (e.g., *"I am at one minute and twelve seconds"* [P3], or *"Go to the marker that is on the second minute of the video"* [P3]). In addition, [P8] used temporal references to inform his partner where to go: *"For the front view [on the structure] come to 2 minutes and 20 seconds, and then go right to zero minutes"*. At a different stage of the activity, P8 told his partner to go to a building, *"At the third minute"* of the video. The adoption of temporal navigation strategy seemed to result in participants having a harder time maintaining awareness of one another in the study video tours.

How participants use tools and how they talk to each other reflects how they are thinking about the video and the environment. Tourgether360 seems to engender and enable the possibility to navigate the video through a spatial mental model, rather than a strictly time-based model.

6.6.4 Use, Persistence, and Ownership of Markers. While participants did not encounter major challenges using the markers to complete the study tasks, we observed that how participants seemed to want to use markers encompassed a broad class of usages that we had not envisioned at the outset. We discuss how groups used the markers to coordinate activity. Then, we discuss the challenges they encountered with the fact that the markers are persistent rather than ephemeral, which leads to clutter. Finally, we discuss the tensions around ownership that occur because of this persistence.

Use of Markers: Detailed Allocentric Coordination. The markers served as an important mean of allocentric navigation, coordination, and communication. Participants used them to mark places that they would return to, mark places that they wanted to talk about, and discuss the architecture under the markers. As an example, we observed how [P3] requested his partner to put a marker to understand the position that the former wanted to discuss: *"Just put a marker, and I'll come there."* Subsequently, the participants discussed the location together and elected it as a location of choice. The markers helped the collaborators to understand each other's perspective when referring to places in the video they deemed interesting. They anchored conversation, where long verbal exchanges tended to occur around and about the markers. For example, the markers helped [P5] to describe his opinion about the location he thought was particularly relevant and interesting: *"Do you see my marker? I think that this one could be the [main shrine in the shrine complex]. This (while referring to the other marker) should be the main gate."*

Participants also used markers as a coordination mechanism to maintain awareness of others' activities when they were out of view—as a form of visual feedthrough. This happened most frequently when participants split apart during loosely coupled parts of the task. They monitored one another's activities through the markers that were visible both in the minimap and main view, which served as a way of illustrating work being done. For example, in one study session we observed how one of the participants [P2] jumped to the end of the video immediately after the session had started and proceeded marking several potentially relevant locations. The pair then had discussed whether these markers are meaningful or not, even though one of the participants [P1] had never visited this part of the video himself and viewed them via the minimap.

Although the markers were fundamentally grounded in the space, participants would sometimes refer to an individual marker with a combination of spatial and temporal speech. Typically, this was to clarify which specific marker they were referring to (if a spot had multiple markers). For example, at the end of the study session, [P1] said to his partner: *"OK, so let's [choose] the markers placed at 35 seconds, 52 seconds, and 3 minutes [as our final choice of relevant locations]"*. Thus, the markers enabled the participants to meaningfully organize their work, dividing the tasks between them. It also served as a powerful support for communication, allowing the participants to ground their discussions while understanding what precisely the other users referred to.

Persistence of Markers: Lost Ephemeral Context. We implemented markers as persistent annotations answering Study 1 participants' desire to bookmark locations. However, their persistency could also lead to communication issues. Participants sometimes had trouble locating specific markers further along in the task once many markers had been created. For example, when [P4] wanted to direct his partner's attention to a specific marker in the environment among a series of markers that were already there, his partner was confused which specific marker was under discussion. After spending roughly 15-20 seconds trying to locate the marker, they gave up and continued reviewing other locations. Thus, while the markers served as opportunities for coordination and discussion, participants still needed ways of clarifying which marker their partners should look at. Beyond this, the markers in of themselves were sometimes insufficient to clarify what partners should look at. For instance, [P2] reports, *"It was mostly verbal descriptions [to clarify]. The markers were not so useful because you cannot see them from everywhere."* Similarly, [P3] reported, *"Even when you see the marker, it is not always clear what exactly it refers to"*. In this respect, the markers served to get partners to roughly the right location, and then the participants would need to clarify what to look at with verbal prompts. We observed all the participants actively referencing the points of interest in the video using verbal expressions, particularly verbal deictic referencing, like: *"look to the right"* or *"go forward and then slow down"* [P9].

The challenge is that while it was easy to create markers, once created, it would take an equal effort to remove them. Thus, markers would be left throughout the video during the study session—regardless of whether they had been placed to just attract someone's attention temporarily as part of conversation (e.g., "Look at this!") versus markers that were intended to mark significant points of interest (e.g., parts of the video that the group expected to return to). The participants explained that because some of the markers have already lost their significance or simply due to many markers crowding the view, they were confused as to what each marker meant, especially toward the end of the study sessions. For example, [P2] stated that after he encountered the markers that were put there by his partner in the experiment, *"It was hard to understand what this marker refers to."* [P2].

Markers served as a monolithic "spatial communication" mechanism, with participants being unable to distinguish their intention, or what they referred to. Several participants suggested being able to color-code markers to signify intention. Others suggested adding the ability to provide verbal annotation with markers to *"describing the context behind the reason for putting this marker"* [P3]. In practice, it may also be useful to include mechanisms that support temporary, ephemeral deictic reference, such as telepointers, as described by [8]. Such a visual mechanism could support conversation without the need to clutter the environment in a persistent way.

Ownership of Markers. We rarely saw markers deleted, regardless of the group. In Tourgether360, markers denote the creator with a label; however, as described above, the reason for a marker's creation was not clear. Participants described feeling reluctant to delete others' markers—even when the environment was extremely cluttered with markers. [P7] explained this as a problem of ownership: *"This marker is not mine, I had not created it, so I don't think that I should have the right to delete it. I don't know why [another person] created it, so I will not meddle with it".* Similarly, P12 mentioned, *"Towards the end of the session, we had marked a lot of points because of which it became confusing. Although I found some of the markers placed by [my partner] to be clumsy, I was reluctant to delete them. Maybe it could be implemented such that if I delete a marker, it gets deleted for me only, so that [my partner] can revisit the marker he had bookmarked."*

Thus, while the markers served a coordinating role for our participants, they still created situations of ambiguity that needed to be resolved verbally. Furthermore, their persistence caused clutter—particularly later on during task sessions—even when their presence was only intended for ephemeral purposes.

7 DISCUSSION AND FUTURE WORK

As ongoing COVID-related restrictions continue to disrupt tourism and travel, social experiences in virtual spaces may become an important way of contributing to people's social and personal well-being. Experiences around 360° videos provide rich, engageable, and realistic content that is well-suited for a range of touristic experiences. However, the current design of 360° video players makes it hard to comfortably enjoy and navigate such videos with others—particularly when the goal is to communicate, coordinate and socializing with other people. Tourgether360 extends prior work on providing spatial means of understanding and watching 360° videos [24] by supporting spatial navigation on an architectural minimap, and simulates a co-habited space with other collaborators. The findings from our study reveal several new questions and issues that are worthy of future study:

Spatial/Semantic Metaphors for Navigating 360° Video. The approach illustrated in Tourgether360 encourages navigation and operation with the data through spatial means. In practice, we observed that references were made to the architecture in the environments (e.g., "Look at the entrance") as well as in relation to participants and partners' avatars (e.g. "Look to my left," or "Look to your left"). In effect, the minimap provides a spatial overview that is akin to the semantic transcripts that can be used to cross-reference into video (e.g. [11]). While it is possible to imagine many semantic layers being applied to videos to support navigation (e.g., labeling buildings; egress; people or cars; transcripts; businesses, etc.), it is interesting to consider what kinds of semantic metaphors are important to effectively navigate video. To some extent, this likely depends on the nature of the task, and the intention for navigating and studying the video in the first place. By studying the mental maps that people develop as they watch and discuss 360° videos, we may be able to uncover the most effective types of labels and metaphors for navigating 360° video.

"Inhabiting Together" rather than "Watching Together." While considerable work has explored remote, social TV watching experiences (e.g. [12]), one of the enduring challenges has been to design experiences that are enjoyable for people to use—as if they were collocated and watching together. The approach that Tourgether360 takes tries to take that a step further—rather than considering 360° videos something that can be watched together, Tourgether360 allows making 360° videos something people can inhabit together: an experience that was made evident by participants' remarks. The flexibility to move around in the video space independently enhanced this sensation of control and immersion. It is interesting to consider what other kinds of media we might be able to design immersive experiences where people feel as if they are "together." We know, for instance, that text conversation with other viewers of livestream broadcasts (and potentially people in the live video) can help people feel "together" (e.g. [12]). One way to conceptualize these conversations is that they are creating virtual conversational places that viewers occupy together [13]. What are other ways that we can create "places" that viewers can cohabit together without creating undesirable burden on the ways they can interact?

Focus+Context as a Metaphor for Collaborative Coordination. In our study, participants move between periods of loosely coupled and tightly coupled collaborative work [10, 11]. To coordinate these shifts, participants used the minimap to understand where their partner was, as well as to develop a quick understanding of what they were looking at, and where they were to go. This "context gathering" step is provided by the minimap through the avatar representation (along with the avatar's orientation cone) and the markers. We noted that participants typically studied this minimap first before teleporting themselves to their destination. This resonates with Schneiderman's adage to support overview before providing details on demand [32]. Our study suggests that beyond simply providing ongoing "workspace awareness" of other's activities, awareness tools can also provide this sort of "context" information for when a collaborator needs to shift their view of the workspace dramatically. It also suggests that when collaborators need to dramatically shift their position in the workspace, doing so smoothly can provide them with some understanding of the target destination (and the work collaborators have done there) is useful.

Future Improvements to Tourgether360. The studies demonstrated that Tourgether360 can still be substantially improved as a collaborative interface. For instance, we observed many nuanced uses of the pointing mechanism, some where participants intended them to be persistent landmarks, and other times when they were clearly intended to support ongoing conversation and so not intended to be persistent. Participants had no way of distinguishing between landmarks—who had created which ones, and for what purpose. In future iterations of Tourgether360, users will be able to customize these artificial landmarks with colour, annotations, and opacity. While the opacity of the landmarks did not create issues for participants in our studies, they do obscure the environment. Participants' robots could also be coloured, which might help with distinguishing between users when there are more than just dyads. Furthermore, the design of the avatars themselves may affect feelings of shared presence. It may be possible, for instance, to explore more humanoid avatars to see how this enhances the feeling of being together within the video space. Finally, Tourgether360 the spatial metaphor of traversing the tour trajectory breaks down somewhat if the original video has moments where the camera operator stopped moving or moves at a different speed; we will explore ways of visualizing these changes in the focus view.

Experiences of Non-Gamers. Many participants in our sample had substantial experiences playing video games; it is worth exploring what the experiences of non-gamers with Tourgether360 would be—specifically, would they be able to easily navigate the environment? Would they still find the pseudo-spatial navigation techniques effective? Many of the gamers had already seen and experienced the kinds of metaphors we rely on in Tourgether360—minimaps, traces on the environment, and indeed navigating 3D environments on a desktop computer. While these techniques have become the interaction vernacular for video games, they are also elements that require some learning. It may be that "acquisition cost" of learning these techniques was paid off by participants long ago, and that non-gamers would need to learn them.

Extensions for Non-Desktop Interfaces. Our test environment focused on the experience of navigating 360° videos from a desktop. Our future work will explore how to extend these interaction techniques for both tablet displays as well as head mounted displays. We expect that the increased immersion into the 3D environment with these interaction techniques will create new opportunities for exploring how to effectively navigate 360° videos.

Extensions for Non-Tour 360° Videos. Our approach relies on 360° video tours, where a single camera moves through a relatively fixed architectural space. Yet, while many videos on popular platforms are recorded as tour videos, not all 360° videos are recorded in this way. Many 360° videos do not have landmarks that can be effectively used for spatial navigation; in these cases, Tourgether360 would not be effective. For instance, some 360° videos are taken from a static position (i.e. they do not move), and so there would be no spatial context that could be accommodated properly (although, it might be interesting to explore how slit scanning could create a 2D spatial context that could be navigated [16]). Similarly, other 360° videos might not have static landmarks that would be effective for spatial navigation–for instance, videos of skydiving, or videos navigating through a crowd of people (where there is very little that is fixed in the scene). We need to understand how non-tour 360° videos are explored and watched to understand what kinds of approaches would be appropriate for viewing collaboratively with others.

Limitations The participants of both our studies were mostly male and young, reflecting the current bias in VR users [26]. Our future work will engage a broader population to understand their respective experiences with Tourgether360.

8 CONCLUSION

Tourgether360 provides a way for collaborators to explore and navigate 360° tour videos together with the metaphor of a shared space. This can be used in combination with existing temporal navigation mechanisms for an effective 360° video exploration. We observed in our studies that users can easily adopt spatial metaphors for navigation, and that additional communication and coordination mechanisms may be necessary for a truly effective experience. In spite of these limitations, participants enjoyed inhabiting and exploring new shared space together. Tourgether360 demonstrates that even simple augmentations of 360° videos can change the nature of collaborative experiences, and sheds new light on how we can further improve such collaborative experiences.

ACKNOWLEDGMENTS

This research was generously supported by the National Science and Engineering Research Council (RGPIN-2017-04883) and the Mitacs Globalink Research Internship program. We thank our participants for their time, and our anonymous reviewers for providing excellent advice on how to improve earlier drafts of this manuscript.

REFERENCES

[1] Steve Benford, John Bowers, Lennart E Fahlén, Chris Greenhalgh, and Dave Snowdon. 1995. User embodiment in collaborative virtual environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 242–249.

[2] Steve Benford and Lennart E Fahlén. 1993. Awareness, focus, and aura: A spatial model of interaction in virtual worlds. *ADVANCES IN HUMAN FACTORS ERGONOMICS* 19 (1993), 693–693.

[3] Samuel Dodson, Ido Roll, Matthew Fong, Dongwook Yoon, Negar M. Harandi, and Sidney Fels. 2018. Active Viewing: A Study of Video Highlighting in the Classroom. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 237–240. https://doi.org/10.1145/3176349.3176889

[4] Nicolas Ducheneaut, Robert J Moore, Lora Oehlberg, James D Thornton, and Eric Nickell. 2008. Social TV: Designing for distributed, sociable television viewing. *Intl. Journal of Human–Computer Interaction* 24, 2 (2008), 136–154.

[5] Jeff Dyck and Carl Gutwin. 2002. Groupspace: A 3D Workspace Supporting User Awareness. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI EA '02)*. Association for Computing Machinery, New York, NY, USA, 502–503. https://doi.org/10.1145/506443.506450

[6] Diana Fonseca and Martin Kraus. 2016. 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* (Tampere, Finland) *(AcademicMindtrek '16)*. Association for Computing Machinery, New York, NY, USA, 287–296. https://doi.org/10.1145/2994310.2994334

[7] Mike Fraser, Steve Benford, Jon Hindmarsh, and Christian Heath. 1999. Supporting Awareness and Interaction through Collaborative Virtual Interfaces. In *Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology* (Asheville, North Carolina, USA) *(UIST '99)*. Association for Computing Machinery, New York, NY, USA, 27–36. https://doi.org/10.1145/320719.322580

[8] S. Greenberg, C. Gutwin, and M. Roseman. 1996. Semantic telepointers for groupware. In *Proceedings Sixth Australian Conference on Computer-Human Interaction*. 54–61. https://doi.org/10.1109/OZCHI.1996.559988

[9] Chris Greenhalgh and Steven Benford. 1995. MASSIVE: A Collaborative Virtual Environment for Teleconferencing. *ACM Trans. Comput.-Hum. Interact.* 2, 3 (sep 1995), 239–261. https://doi.org/10.1145/210079.210088

[10] Carl Gutwin and Saul Greenberg. 1998. Design for individuals, design for groups: tradeoffs between power and workspace awareness. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*. 207–216.

[11] Carl Gutwin and Saul Greenberg. 2002. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)* 11, 3 (2002), 411–446.

[12] William A. Hamilton, John Tang, Gina Venolia, Kori Inkpen, Jakob Zillner, and Derek Huang. 2016. Rivulet: Exploring Participation in Live Events through Multi-Stream Experiences. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (Chicago, Illinois, USA) *(TVX '16)*. Association for Computing Machinery, New York, NY, USA, 31–42. https://doi.org/10.1145/2932206.2932211

[13] Steve Harrison and Paul Dourish. 1996. Re-place-ing space: the roles of place and space in collaborative systems. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. 67–76.

[14] Kyoungkook Kang and Sunghyun Cho. 2019. Interactive and Automatic Navigation for 360° Video Playback. *ACM Trans. Graph.* 38, 4, Article 108 (jul 2019), 11 pages. https://doi.org/10.1145/3306346.3323046

[15] Jie Li, Yiping Kong, Thomas Röggla, Francesca De Simone, Swamy Ananthanarayan, Huib de Ridder, Abdallah El Ali, and Pablo Cesar. 2019. Measuring and Understanding Photo Sharing Experiences in Social Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300897

[16] Jiannan Li, Jiahe Lyu, Mauricio Sousa, Ravin Balakrishnan, Anthony Tang, and Tovi Grossman. 2021. Route Tapestries: Navigating 360° Virtual Tour Videos Using Slit-Scan Visualizations. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 223–238. https://doi.org/10.1145/3472749.3474746

[17] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. 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* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2535–2545. https://doi.org/10.1145/3025453.3025757

[18] Ville Mäkelä, Tuuli Keskinen, John Mäkelä, Pekka Kallioniemi, Jussi Karhu, Kimmo Ronkainen, Alisa Burova, Jaakko Hakulinen, and Markku Turunen. 2019. What Are Others Looking at? Exploring 360° Videos on HMDs with Visual Cues about Other Viewers. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video* (Salford (Manchester), United Kingdom) *(TVX '19)*. Association for Computing Machinery, New York, NY, USA, 13–24. https://doi.org/10.1145/3317697.3323351

[19] Raúl Mur-Artal and Juan D. Tardós. 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Transactions on Robotics* 33, 5 (2017), 1255–1262. https://doi.org/10.1109/TRO.2017.2705103

[20] Alaeddin Nassani, Li Zhang, Huidong Bai, and Mark Billinghurst. 2021. ShowMeAround: Giving Virtual Tours Using Live 360 Video. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 168, 4 pages.

https://doi.org/10.1145/3411763.3451555

[21] Luís A. R. Neng and Teresa Chambel. 2010. Get around 360° Hypervideo. In *Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments* (Tampere, Finland) *(MindTrek '10)*. Association for Computing Machinery, New York, NY, USA, 119–122. https://doi.org/10.1145/1930488.1930512

[22] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. CollaVR: Collaborative In-Headset Review for VR Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 267–277. https://doi.org/10.1145/3126594.3126659

[23] Cuong Nguyen, Stephen DiVerdi, Aaron Hertzmann, and Feng Liu. 2017. Vremiere: In-Headset Virtual Reality Video Editing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5428–5438. https://doi.org/10.1145/3025453.3025675

[24] Gonçalo Noronha, Carlos Álvares, and Teresa Chambel. 2012. Sight Surfers: 360º Videos and Maps Navigation. In *Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia* (Nara, Japan) *(GeoMM '12)*. Association for Computing Machinery, New York, NY, USA, 19–22. https://doi.org/10.1145/2390790.2390798

[25] Amy Pavel, Björn Hartmann, and Maneesh Agrawala. 2017. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 289–297. https://doi.org/10.1145/3126594.3126636

[26] Tabitha C. Peck, Laura E. Sockol, and Sarah M. Hancock. 2020. Mind the Gap: The Underrepresentation of Female Participants and Authors in Virtual Reality Research. *IEEE Transactions on Visualization and Computer Graphics* 26, 5 (2020), 1945–1954. https://doi.org/10.1109/TVCG.2020.2973498

[27] Benjamin Petry and Jochen Huber. 2015. Towards Effective Interaction with Omnidirectional Videos Using Immersive Virtual Reality Headsets. In *Proceedings of the 6th Augmented Human International Conference* (Singapore, Singapore) *(AH '15)*. Association for Computing Machinery, New York, NY, USA, 217–218. https://doi.org/10.1145/2735711.2735785

[28] Thammathip Piumsomboon, Gun A. Lee, Jonathon D. Hart, Barrett Ens, Robert W. Lindeman, Bruce H. Thomas, and Mark Billinghurst. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173620

[29] Sylvia Rothe, Mario Montagud, Christian Mai, Daniel Buschek, and Heinrich Hußmann. 2018. Social viewing in cinematic virtual reality: Challenges and opportunities. In *International Conference on Interactive Digital Storytelling*. Springer, 338–342.

[30] Gustavo Alberto Rovelo Ruiz, Davy Vanacken, Kris Luyten, Francisco Abad, and Emilio Camahort. 2014. Multi-Viewer Gesture-Based Interaction for Omni-Directional Video. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4077–4086. https://doi.org/10.1145/2556288.2557113

[31] Mehrnaz Sabet, Mania Orand, and David W. McDonald. 2021. Designing Telepresence Drones to Support Synchronous, Mid-Air Remote Collaboration: An Exploratory Study. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 450, 17 pages. https://doi.org/10.1145/3411764.3445041

[32] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.

[33] Anthony Tang and Omid Fakourfar. 2017. Watching 360° Videos Together. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4501–4506. https://doi.org/10.1145/3025453.3025519

[34] Anthony Tang, Omid Fakourfar, Carman Neustaedter, and Scott Bateman. 2017. Collaboration with 360° Videochat: Challenges and Opportunities. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 1327–1339. https://doi.org/10.1145/3064663.3064707

[35] Audrey Tse, Charlene Jennett, Joanne Moore, Zillah Watson, Jacob Rigby, and Anna L. Cox. 2017. Was I There? Impact of Platform and Headphones on 360 Video Immersion. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2967–2974. https://doi.org/10.1145/3027063.3053225

[36] Cheng Yao Wang, Mose Sakashita, Upol Ehsan, Jingjin Li, and Andrea Stevenson Won. 2020. *Again, Together: Socially Reliving Virtual Reality Experiences When Separated.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376642

[37] Nelson Wong and Carl Gutwin. 2010. Where Are You Pointing? The Accuracy of Deictic Pointing in CVEs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI ’10)*. Association for Computing Machinery, New York, NY, USA, 1029–1038. https://doi.org/10.1145/1753326.1753480

[38] Nelson Wong and Carl Gutwin. 2014. Support for Deictic Pointing in CVEs: Still Fragmented after All These Years’. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW ’14)*. Association for Computing Machinery, New York, NY, USA, 1377–1387. https://doi.org/10.1145/2531602.2531691

[39] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. 2020. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing* 14, 1 (2020), 5–26.

[40] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S. Fels. 2019. "Can You Believe [1:21]?!": Content and Time-Based Reference Patterns in Video Comments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI ’19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300719