Predict-and-Drive: Avatar Motion Adaption in Room-Scale Augmented Reality Telepresence with Heterogeneous Spaces

Xuanyu Wang, Hui Ye, Christian Sandor, Weizhan Zhang, Hongbo Fu

Predict-and-Drive: Avatar Motion Adaption in Room-Scale Augmented Reality Telepresence with Heterogeneous Spaces

Xuanyu Wang, Hui Ye, Christian Sandor, Weizhan Zhang, and Hongbo Fu

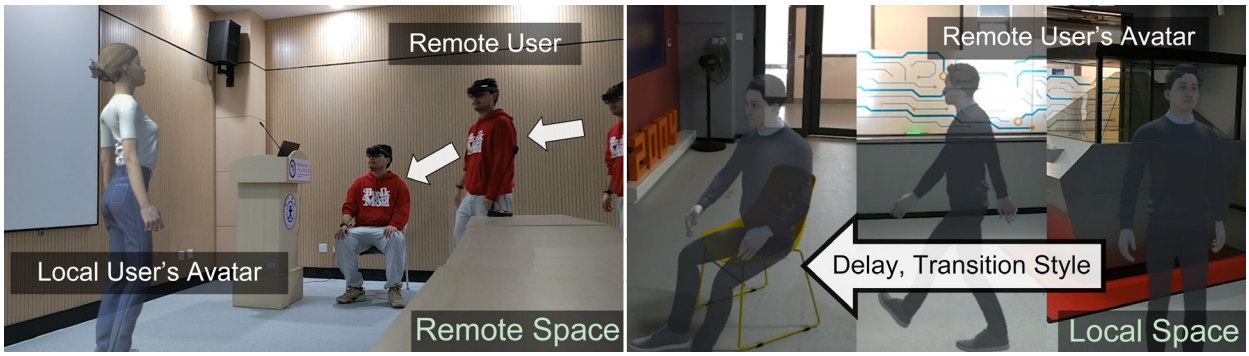

Fig. 1. Two users separated in their respective remote heterogeneous physical spaces can move and interact with each other via their corresponding avatars. The right scene is seen from the Local User's perspective, while the left scene is from a third-person perspective. In the local space, the Local User (not shown in this figure) can interact with the Remote User's avatar (Right), which is a representation of the teleported remote user. The avatar transitions adaptively with respect to the Remote User. Symmetrically in the remote space (Left), the physical user (i.e., the man in this example) interacts with the Local User's avatar simultaneously.

Abstract— Avatar-mediated symmetric Augmented Reality (AR) telepresence has emerged with the ability to empower users located in different remote spaces to interact with each other in 3D through avatars. However, different spaces have heterogeneous structures and features, which bring difficulties in synchronizing avatar motions with real user motions and adapting avatar motions to local scenes. To overcome these issues, existing methods generate mutual movable spaces or retarget the placement of avatars. However, these methods limit the telepresence experience in a small sub-area space, fix the positions of users and avatars, or adjust the beginning/ending positions of avatars without presenting smooth transitions. Moreover, the delay between the avatar retargeting and users' real transitions can break the semantic synchronization between users' verbal conversation and perceived avatar motion. In this paper, we first examine the impact of the aforementioned transition delay and explore the preferred transition style with the existence of such delay through user studies. With the results showing a significant negative effect of avatar transition delay and providing the design choice of the transition style, we propose a *Predict-and-Drive* controller to diminish the delay and present the smooth transition of the telepresence avatar. We also introduce a grouping component as an upgrade to immediately calculate a coarse virtual target once the user initiates a transition, which could further eliminate the avatar transition delay. Once having the coarse virtual target or an exactly predicted target, we find the corresponding target for the avatar according to the pre-constructed mapping of objects of interest between two spaces. The avatar control component maintains an artificial potential field of the space and drives the avatar towards the target while respecting the obstacles in the physical environment. We further conduct ablation studies to evaluate the effectiveness of our proposed components.

Index Terms—AR Telepresence, Avatar Motion Adaption, Heterogeneous Spaces, Redirected Walking

1 Introduction

Avatar-mediated symmetric Augmented Reality (AR) telepresence gives users the illusion of human teleportation. It enables spatially separated users to interact with each other through their avatars overlaid on the local physical space. Due to the significant improvement in the level of co-presence, such techniques are continuously gaining traction both in research communities [21, 23] and industries[1].

The Holoportation system [23] well demonstrates such AR telepresence with impressive real-time spatial audio and 3D visual effect. However, requiring two almost identical spaces significantly limits the application scenarios of this system, since most of our spaces are different in shapes and installations. Simply copying a user's motion to his/her avatar in a remote space could easily lead to anomalies such as the avatar penetrating obstacles and interacting with air. To tackle this problem, researchers propose to generate and use a mutual movable space out of the remote spaces involved in the experience [11, 13, 19], but such methods restrict user interactions to a small sub-area of the space. Other approaches introduce avatar position retargeting in two spaces by designating static correspondence in remote and local spaces at the beginning [21, 24]. But most of them disable the free movement of users.

- *Corresponding authors: Hongbo Fu and Weizhan Zhang.*
- *Xuanyu Wang is with MOEKLINNS Lab, School of Computer Science and Technology, Xi'an Jiaotong University and the School of Creative Media, City University of Hong Kong. E-mail: xwang2247-c@my.cityu.edu.hk.*
- *Hui Ye and Hongbo Fu are with the School of Creative Media, City University of Hong Kong. E-mail: {hui.ye, hongbofu}@cityu.edu.hk.*
- *Christian Sandor is with Université Paris-Saclay / CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN). E-mail: chris.sandor@gmail.com.*
- *Weizhan Zhang is with MOEKLINNS Lab, School of Computer Science and Technology, Xi'an Jiaotong University. E-mail: zhangwzh@xjtu.edu.cn.*

Manuscript received 11 March 2022; revised 11 June 2022; accepted 2 July 2022. Date of publication 01 September 2022; date of current version 03 October 2022. Digital Object Identifier no. 10.1109/TVCG.2022.3203109

[1] https://www.microsoft.com/mesh

Yoon et al. [40] propose a learning-based approach to constantly retarget the placement of an avatar to a position that has similar semantics to the corresponding remote user's position during the telepresence experience. However, this method retargets the avatar's placement when the user has already arrived at a new target position, leaving the avatar delayed at the original point during the user's transition period. It results in a sudden jump instead of a smooth transition when retargeting the avatar's position, and thus can cause the motion and conversation context to be out of sync when taking the semantics of the real-time verbal conversation into account.

For simplicity, we focus on the local space with one local user (simply referred as "the user" when there is no ambiguity) and one remote user's avatar (referred as "the avatar") since the experience is symmetric and can be scaled to multi-user cases. As illustrated in Figure 1, the user can interact with the avatar representing the teleported remote user both verbally and non-verbally. The above-mentioned issues lead to our first question: in such room-scale avatar-mediated AR telepresence, when the remote user initiates a locomotion transition in the remote space, **how do the existence of the transition delay and the lack of a transition process on the avatar affect the user's perception of its locomotion (RQ1)?** To answer this question, we conduct our first pilot user study to examine this impact with respect to the user's rating on the degree of similarity to the anticipated transition, the semantic rightness considering the conversation context, and the overall preference. The results show that the transition delay and the instant change in the placement of the avatar can have significant negative influences on the user's experience. This motivates us to explore solutions to diminish such transition delay and instant change for more natural and smooth interactions. Specifically, 1) a prediction of the remote user's target position, which is mapped to the local space as the avatar's target, is necessary to drive the avatar in advance to diminish its transition delay, and 2) the process of the avatar's transition from place to place should be clearly presented to the user. Moreover, our first study also verifies that 3) the basic rule of obstacle avoidance when presenting telepresence avatars is still important for a desired transition of the avatar.

One possible solution to predict the remote user's target position or future path is to leverage the methods in Redirected Walking (RDW) controllers [3, 27, 41]. However, due to the error-sensitive and dynamic nature of the AR telepresence and the lack of a history trajectory, these methods cannot be directly adopted in our avatar-mediated AR telepresence scenario. Moreover, due to the intrinsic ambiguity of the remote user's motion and conversation context, accurate real-time target prediction with zero delay in potentially dynamic AR spaces is hard to achieve in most of the time. Subsequently, our next question arises: **what would be the best transition style for the avatar when there exists a relatively small delay before the exact target of the remote user can be predicted (RQ2)?** To answer this question, we compare three possible transition styles through our second pilot user study. The results show that adjusting the transition speed is the most preferred style regarding the same three metrics used in our first user study. We distill these findings as the guidance and design choice for further studies.

With the results from the two pilot studies, we observe that, from the locomotion perspective, the controller for AR telepresence should be able to predict the remote user's target and generate proper avatar transitions adaptively. Therefore, we propose a novel conceptual solution paradigm for avatar motion retargeting that comprises a (remote) user target prediction component and an avatar transition control component. It is informed by our observations of how our problem is related to the RDW problem in VR, following the mentality of Williams et al. [34]. We then implement the proposed paradigm into a proof-of-concept prototype. We adopt the notion of visibility polygon and the artificial potential field method [30] from RDW controllers to implement the target prediction and avatar control components in our prototype, respectively. In order to further eliminate the transition delay, we propose a grouping component as an upgrade of the target prediction. We initially evaluate the effectiveness of our proposed method by comparing the trajectories generated by the prototype (with an ablation of the grouping component) with the corresponding real human moving trajectory through a user study, and visually demonstrating the delay reduction by the grouping component in simulation environments.

The main contributions of this work are threefold.

- Through two pilot studies, we identify the problem of avatar transition delay in AR telepresence, explore its impact, and find the most preferred avatar transition style when this delay is inevitable.
- Informed by the observed similarities and differences between the AR telepresence problem and the VR RDW problem, we propose a novel conceptual paradigm that decomposes the motion retargeting of the telepresence avatar into two coupled tasks (target prediction and avatar control), providing a new perspective and facilitating new potential solutions.
- We introduce the Predict-and-Drive avatar motion adaption controller as a proof-of-concept implementation. The initial evaluation through a user study proves the necessity and effectiveness of each proposed component.

2 Related Work

2.1 Heterogeneous Space in AR Telepresence

Early AR telepresence endeavours realize the notion of teleporting a remote peer into a local physical space through real-time 3D capturing and reconstruction. They teleport a user to sit on a local sofa using spatial projection [24], or to sit near a table with the right occlusion using AR Head Mounted Display (HMD) [21]. However, they do not support free movement after the remote user is "teleported". Holoportation [23] enables extremely high-quality teleportation and free movement afterwards by configuring the spaces identical to align the virtual and physical contents and thus making it possible to present the exact copy of each other.

The emerging consumer-grade AR and VR HMDs bring such telepresence experience into our daily life. Instead of pursuing high-quality capturing and reconstruction, systems such as Microsoft Mesh and Horizon Workrooms [2] use animated virtual avatars to represent the teleported users. Getting rid of the specifically and expensively configured spaces is necessary to realizing the telepresence in everyday life, as described by Yoon et al. [40]. But the heterogeneity of spaces comes as a new problem with broader using scenarios. In this circumstance, simply copying a user's motion to his/her corresponding avatar in a remote, different space would cause unnatural interactions between the avatar and the physical space, and convey confusing semantics.

In order to prevent the avatar from penetrating obstacles, researchers propose various methods to generate a mutual obstacle-free space. For example, they propose to create a consensus reality by aligning two spaces with regard to the optimization of several geometry factors [19] and giving advice on altering the space to further expand the virtual common ground [11]. Kim et al. [13] stretch the space along the x-axis and y-axis separately to achieve a larger overlay area. Such mutual spaces constrain the user in a small area of a room, which could get smaller or even disappear with the spaces being less similar. Moreover, it fails to present the user's interaction with physical objects, e.g., sitting on a local chair, which is out of the mutual space. Our method aims to present room-scale experiences where a user's daily activities can all be adaptively presented in another space according to the physical installation.

Yoon et al. [40] share our goal to present the telepresence in the whole space. They propose a learning-based approach to retarget an avatar to a new position that is the most similar to the corresponding remote user's position with regard to the interaction, pose accommodation, and spatial semantics. Such an avatar positioning approach based on the scene semantics has been studied in the literature [16]. Jo et al. [10] also determine the avatar's position according to the rigid transformation calculated by an object-to-object mapping. However, these methods can only find the right position for the avatar given a user's position. When implementing these methods in real-time telepresence,

[2] https://www.oculus.com/workrooms

it will lead to a considerable delay since the user's position can only be settled when completing the whole transition. Moreover, they do not present a natural transition process of the avatar corresponding to the change of position, which is one of the main subjects of our work.

2.2 Avatar Motion Adaption

In order to present natural avatar motions and transitions while respecting the environment installations, existing methods often assume a script of what actions virtual avatars will perform. Tahara et al. [29] extract the space and an avatar's action into scene graphs. They manage to generate AR contents with natural and semantically right avatar movement that adapts to the physical environment by matching the two graphs. Huang et al. [8] introduce a motion planning framework to generate life-like demonstrator movements for virtual avatars given an observer's position and a demonstration task. It guarantees the avatar's visibility from the observer, and the motion semantic rightness, e.g., the ability to avoid obstacles, reach an exact target, and maintain gaze behaviours. Specifically for the sitting motion, Kim et al. [15] introduce a method to match key points on joint trajectories to fit an avatar's motion into different object shapes.

The methods mentioned above can adapt an avatar's motion according to physical spaces to generate a natural and semantically right movement presentation. However, they are not suitable for our AR telepresence scenario for 1) the lack of real-time performance and 2) the requirement of predefined task descriptions, which are unavailable since, ideally, users should be allowed to act freely in AR telepresence.

2.3 Agent Control in Redirected Walking

Researchers have proposed plenty of works about RDW algorithms. Thomas et al. [30] use an artificial potential field as the reference to calculate RDW gains. Williams et al. [34] leverage a visibility polygon to redirect the user to an area in the Physical Environment (PE) that is similar in alignment to the current position in the Virtual Environment (VE). There are also predictive methods to include passive haptic feedback to further improve the immersion [3, 27, 41]. The simple ones based on movement are highly error-prone. They work well because RDW controllers are highly error-tolerant since the manipulation is unperceivable. It does not matter if several wrong predictions occur first as long as the right prediction lasts long enough. The complex ones need the information of the user's history trajectory and can only deal with static scenes. Avatar-mediated AR telepresence provides a dynamic and real-time experience, making it hard to directly apply existing RDW approaches to retargeting avatar motion in AR telepresence.

2.4 Perception of Avatar Locomotion and Behaviour

There exist a wide range of studies investigating how a user's third-person-view perception is affected by various factors of an avatar in AR telepresence and virtual 3D experience, e.g., appearance [1, 38], size [33], diegetic representation [6], and realism [17]. Especially for the avatar locomotion and behaviour, Choi et al. [4] examine the effect of an avatar's walking style on the perceived naturalness, similarity to a real user's motion, and the level of intention preservation when varying the walking distance. Kim et al. [14] investigate how the plausibility of an avatar's behaviour affects a user's sense of co-presence. They examine several implausible avatar behaviours, including passing through the door and obstacles without asking for help and not being properly occluded by physical objects.

While most of these previous works examine the visual factors of avatars, our work considers the whole audio-visual AR telepresence experience, taking the semantic synchronization of conversation and avatar motion into account. Existing studies in telepresence examining the asynchronization focus mainly on teleoperation scenarios [7, 22], with the delay caused by end-to-end signal transmission. In VR, there are studies focusing on the effect of asynchronization between a user's motion and his/her self-representing avatar in a virtual environment [5, 18, 25]. However, they focus on the self-embodied first-person-perspective avatar instead of the third-person-point-of-view AR-telepresence avatar of another remote user.

3 Pilot Studies

In this section, we will introduce our pilot studies on the impact of the avatar transition delay and the preferred avatar transition style when such delay is inevitable. Note that what we discuss is the delay introduced by the controller when determining the remote user's target, rather than hardware delays (e.g., network fluctuations, signal transmission, and rendering). We choose a full-body life-like avatar in the studies, since it has been proved to provide the best perception to users in AR remote collaboration by Yoon et al. [38].

3.1 Study 1

As mentioned previously in this paper, retargeting an avatar's position according to its corresponding remote user's position enables room-scale AR telepresence with the right avatar placement semantics, while the problem of the avatar transition delay in this process remains unexplored. To this end, we raise RQ1 (illustrated in Sect. 1) and conduct this first pilot study to answer it. We study the experience in the primary space (the local space) since the whole experience is symmetric.

3.1.1 Participants

We recruited 16 graduate students (13 males and 3 females; average age: 23.8 (SD = 1.06)) from the local campus as the participants. Most of them had a moderate level of AR/VR experience: 4 had no prior AR/VR experience, 10 had experienced several times, 1 had used HMDs extensively, and 1 is an AR/VR developer. The sample size (along with that in the subsequent Pilot Study 2 and Initial Evaluation) is in line with the suggestion (8 to 12) by relevant research in HCI studies [2], and is thus sufficient for drawing our conclusions. We recruited young college students as participants since they are sensitive to XR technology and are our main target users. It was also partially due to a strict COVID-19 campus lockdown during the experiments.

3.1.2 Study Scenarios

We set two daily telepresence scenarios in the study, both with real-time verbal communications and transitions from one place to another. These two scenarios respectively cover the living and study/official scenes and involve common physical objects and interactions for multiple tasks. Scenario 1 involves two users walking to a mutual destination (e.g., a sofa) together, and Scenario 2 involves a user walking towards the other from a distance.

Scenario 1. In the dining room, the local user just finished the dinner with the remote user through experiencing co-presence with his/her avatar. The remote user asks the local user, "How was the dinner?" The local user answers, "It was nice!" Then the remote user asks, "How about we watch some show?" The local user answers, "OK! Let's do it." The remote user says, "Let's go." And then, the user in the remote space immediately starts walking to the living room. When arriving at the sofa in the remote space, the remote user says, "Let me turn on the TV." While sitting down, the remote user turns on the TV simultaneously. The TVs in the two spaces are connected through Internet of Things (IoT), meaning the local TV will be turned on at the same time and play the same content as that in the remote space. With the remote user being invisible, the local user would perceive the above actions of talking and turning on the TV to be initiated by the avatar, and observe the avatar's transition from the kitchen to the sofa in the local living room.

Scenario 2. The user is looking at a digital whiteboard while the avatar is in the living room with a certain distance. The user wants to call the remote user to see the information on the whiteboard and thus asks by shouting at the avatar, "Hey! Come check this board." The remote user hears and answers, "OK, I'm coming", and immediately starts walking from the living room to the local user's avatar and the whiteboard in the remote space, which is connected with the counterpart in the local space through IoT as well. When arriving in front of the remote whiteboard, the remote user says to the local user, "All right, show me." Similarly, the user would feel as calling and talking to the avatar, and observe its transition from the living room to near the whiteboard.

Table 1. Summary of cases in Study 1.

Transition Delay	Transition Process	Delay Style	Obstacle Avoidance
Maximum Delay	Flash	Idle	/
Maximum Delay	Flash	WIP	/
0 Delay	Walk	/	Avoid
0 Delay	Walk	/	No Avoid

3.1.3 Experiment Setup

We use the Microsoft HoloLens 2 to present the AR experience. The avatar and the animation clips are from Mixamo[3]. We construct a virtual copy of the local space in advance, where virtual boxes are aligned with the main physical obstacles, giving an effect similar to the 3D bounding box. This configuration enables proper occlusions for virtual contents since the edges of the virtual and physical objects are roughly aligned, and can provide a simple understanding of the scene semantics. We will introduce its detail in Sect. 4.1. The avatar's part of the conversation is pre-recorded and can be remotely triggered by the researcher during the experiment. We also pre-record the avatar's path from the starting point to the target point from a real user for each of the scenarios. This can make sure the path is semantically right. The configuration of the experiment space is shown in Fig. 2.

3.1.4 Study Cases

We consider four factors (i.e., Transition Delay, Transition Process, Delay Style, and Obstacle Avoidance) in this study. First, regarding the Transition Delay, we consider only two extreme cases, i.e., **Maximum Delay** and **0 Delay**. In the maximum-delay case, the target position for the avatar is set at the end of the transition when the remote user has already arrived at the destination, as in the previous work [40]. We set the maximum delay to 8s since it is the approximate time a real human walks from the dining table to the sofa/whiteboard in our experiment space. We assume that such transitions containing clear changes in interaction status and semantics will usually take a while from a few seconds to more than 10s (as also demonstrated in the previous similar work [40]), depending on the room layout and size. The 0 Delay case represents the ideal situation where the system can predict his/her destination immediately when the remote user starts to move. Second, for the Transition Process, we consider the **Flash** (the avatar's position changes instantly from the starting point to the target point, similar to the VR teleportation) and **Walk** (the avatar smoothly transitions from the starting point to the target point by walking) cases. This variable is dependent on the Transition Delay, where the Maximum Delay corresponds to the Flash transition, and the 0 Delay corresponds to the Walk transition. We also examine the Delay Style for the Maximum Delay case, considering the **Idle** and **Walk In Place (WIP)** cases. Being Idle during the delay, the avatar only stands still in the idle state, while WIP gives the avatar a motion of walking in place. WIP can deliver the semantics of moving, while being Idle is more human-like since human users seldom walk in place. Lastly, we examine the Obstacle Avoidance feature with the **Avoid** and **No Avoid** cases for the 0 Delay case to verify that the obstacle avoidance is still necessary for the telepresence avatar transition. All four cases are shown in Table 1.

3.1.5 Metrics

We refer to similar studies [4] and identify three metrics, namely "similarity to the anticipation", "semantic rightness", and "overall preference". The "similarity to the anticipation" is defined as how much alike the avatar's transition is to the one that a participant anticipates. The "semantic rightness" describes to what extent the conversation context matches the avatar's motion. Lastly, the "overall preference" indicates the subjective ranking of all the tasks. We present the three metrics to the participants using a seven-point Likert scale ranging from 1 to 7. 1 and 7 represent the least and most "similar to the anticipation", "semantically right", and "preferred", respectively.

[3] https://www.mixamo.com/

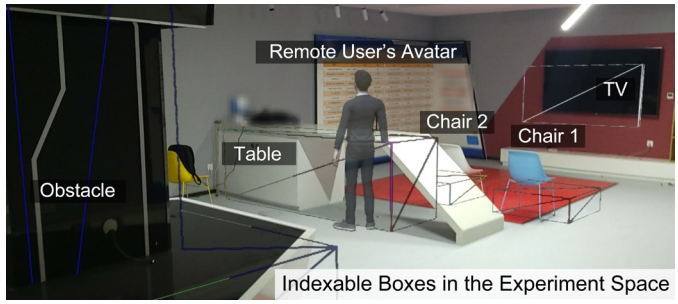

Fig. 2. The configuration of the experiment space (local space) seen from the local user's perspective through the HoloLens. Virtual boxes (with their edges visible to the local user) are aligned with physical objects (i.e., the TV, chairs 1 and 2, table, and the obstacle). The whiteboard in Scenario 2 is out of the current view. The avatar representing the teleported remote user is overlaid in the space.

We make the following hypotheses:

- **H1**: Walking is proved to be effective in increasing the sense of presence compared to other instance position changing methods in VR navigation [31]. We assume analogically in our case, 0 Delay with transition will have significantly higher scores than Maximum Delay with no transition process regarding all three metrics.
- **H2**: WIP will have a significantly higher score in the semantic rightness than Idle, as it is used and claimed effective in previous methods [39, 40].
- **H3**: Avoiding obstacles will lead to significantly higher scores than no handling of obstacle avoidance.

3.1.6 Procedure

First, we introduce the basic concept of the symmetric AR telepresence to each participant through a video tutorial and verbal instructions. Second, after the concept is all clear to a participant, we introduce the two scenarios mentioned above, and ask him/her to play the role of the local user. Then we walk the participant through all the events that will occur during a task in the experience verbally. Afterwards, we clarify the definitions of the three metrics to each participant, and let him/her explore the AR scene for a few minutes wearing the HoloLens device. After the preparation, we start the experiment process.

In the experiment, there will be 8 tasks in total for each participant to experience, 4 tasks in each of the 2 scenarios. The sequence of the 4 cases (corresponding to the 4 tasks in each scenario) shown to the participants is counterbalanced in a hierarchical fashion. We ask each participant to rate their experience after each task using the seven-point Likert scale introduced earlier. They are reminded of the definitions of the three metrics and the meanings of the scores (as in Sect. 3.1.5) and are asked to select the corresponding score under the title of each metric. They can change the score or ask for a replay of any task at any time. After finishing all 8 tasks, we have a small interview for each participant to explore the potential insights and suggestions.

3.1.7 Results

To analyze the study results, we first ran normality tests on the data, and found non-normality distributions. Therefore, we conducted Kruskal-Wallis H-test and Post-hoc Dunn's Test on the data combined from the two scenarios to compare the subjective ratings of all the metrics. We will elaborate on the specific results under the confidence interval of $p < 0.05$ below. The statistical results for the combined data from the two scenarios are shown in Fig. 3.

Similarity. Kruskal-Wallis H-test shows a significant effect of the four conditions (H *Statistic* = 87.91, p = 6.2e-19). We further conducted Dunn's Test for pairwise comparisons and found that Flash-Idle

(M = 2.13, SD = 1.31) and Flash-WIP (M = 2.88, SD = 1.10) have no significant difference (p = 0.095), both having a low score. Walk-Avoid (M = 6.34, SD = 0.75) is significantly better than any other case: Flash-Idle (p = 2.85e-18), Flash-WIP (p = 1.81e-12), and Walk-No Avoid (p = 5.69e-5). Walk-No Avoid (M = 4.22, SD = 0.97) is significantly different from both Flash-Idle (p = 2.7e-6) and Flash-WIP (p = 0.003).

Semantic Rightness. Kruskal-Wallis H-test shows a significant effect of the four conditions (H *Statistic* = 89.05, p = 3.5e-19). According to Dunn's Test, Flash-WIP (M = 2.91, SD = 1.12) has no significant difference (p = 0.084) from Flash-Idle (M = 2.09, SD = 1.17). Walk-Avoid (M = 6.28, SD = 0.77) provides significant more semantic rightness than any other case: Flash-Idle (p = 1.5e-18), Flash-WIP (p = 1.6e-12), and Walk-No Avoid (p = 5.3e-5). Walk-No Avoid (M = 4.25, SD = 1.02) is significantly more right in semantics than both Flash-Idle (p = 2.1e-6) and Flash-WIP (p = 0.003).

Preference. Kruskal-Wallis H-test shows a significant effect of the four conditions (H *Statistic* = 82.14, p = 1.1e-17). We further ran Dunn's Test, and found that Flash-Idle (M = 2.34, SD = 1.36) and Flash-WIP (M = 2.94, SD = 1.29) have no significant difference (p = 0.224). Walk-Avoid (M = 6.16, SD = 0.88) is significantly better than any other case: Flash-Idle (p = 9.0e-17), Flah-WIP (p = 1.2e-12), and Walk-No Avoid (p = 3.1e-5). Flash-Idle and Walk-No Avoid (M = 4.25, SD = 0.88) (p = 3.3e-5), Flash-WIP and Walk-No Avoid (p = 0.003), both have significant differences.

Discussion. **H1 is supported:** The "Walk-Avoid" avatar transition style (i.e., with natural walking and obstacle avoidance) presents the participants a significantly better experience than any other combination regarding all the three metrics. This should be the ideal transition process, which we distill as the ground-truth effect for a telepresence avatar. "Walk-No Avoid" is significantly better than "Flash-Idle" and "Flash-WIP" regarding all the three metrics, indicating that the maximum delay and the resulting lack of transition significantly harm the user's experience even more than the avatar's penetration of obstacles. **H2 is rejected:** We can see, with the maximum delay, "WIP" and "Idle" both score low in all the metrics and have no significant difference. The existence of the delay has the main effect. We observe that in Scenario 1, the participants would be more likely to start to walk to the sofa in the "WIP" condition. In contrast, in the "Idle" condition, they tend to stand still at the starting point until the avatar flashes to the sofa after the delay. This can partially prove the movement indication by the "WIP" motion, but is not reflected in the scores. Some participants feel "the avatar seems stuck in front of the table" in "WIP". This is because the avatar's target direction is not determined during the delay, and thus is walking in place probably facing a wrong direction due to the heterogeneity of the two spaces. Following the previous method [39,40], in Study 2 and the subsequent implementation, we still choose "WIP" for the avatar when the delay problem occurs because of its higher mean scores, though not significant. **H3 is supported:** With transition presented, "Avoid" is significantly better than "No Avoid". Steering the avatar to walk around obstacles is also an important factor in the experience. Besides performing data analysis on the combined data, we have also done the analysis of the data for Scenarios 1 and 2, separately, and found the above conclusions still hold. Please refer to the supplemental document for more details.

3.2 Study 2

The results of Study 1 suggest that we should eliminate the transition delay and try to present a smooth transition process instead of dramatically changing an avatar's position. The purpose of our second user study is to answer the second question (RQ2) raised previously focusing on the transition style in the circumstance with an inevitable small delay. Study 2 shares the same scenarios, experiment setup, metrics, and procedure with Study 1. We set the delay to half of the maximum delay used in Study 1 for the cases in Study 2, i.e., 4s. We also fix the delay style to WIP regarding the observation in Study 1.

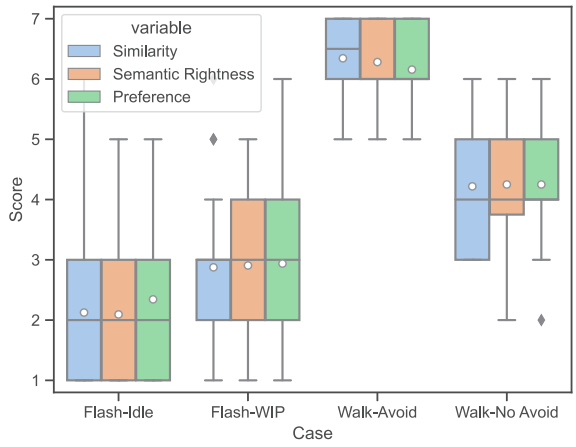

Fig. 3. The statistical results for the combined data from Scenarios 1 and 2 of Study 1.

3.2.1 Participants

We recruited 12 graduate students (8 males, 4 females; average age: 23.6 (SD = 1.08)) from the local campus as the participants. Most of them had a moderate level of AR/VR experience: 3 had no prior AR/VR experience, 8 had experienced several times, and 1 is an AR/VR developer. Due to the COVID-19 restrictions, we had to invite several participants of Study 1 to join Study 2 (and the subsequent Initial Evaluation as well). We believe there is no significant bias since there is no overlap in the independent variables across the studies, which are clearly distinct and were designed and conducted progressively. We provide detailed relevant information in the supplemental material.

3.2.2 Study Cases

We consider three transition styles namely **Const**, **Speed**, and **Hybrid,** in this study. First, for the Const case, the avatar walks towards the target at the original speed (set to 0.6 m/s in this study). Notice that, in this case, the prediction delay can cause a corresponding arrival delay, which would lead to a gap between the originally synchronized conversation context and the remote user's actual transition. Therefore, the potential demand of diminishing the arrival delay leads to the Speed transition style. It speeds up the avatar's transition accordingly (doubled in this case) in order to reach the local target at the same time with the remote user reaching the remote target. In the experiment, we double both the loading step of the avatar path and the animation speed. With the frame rate fixed, both the avatar's translation and animation speeds are doubled with an unchanged stride length (proved to be more preferred by Choi et al. [4]). Third, we can adopt the Hybrid approach to first let the avatar walk at its original speed, and then jump to the local target instantly when the remote user reaches the remote target to synchronize the position and conversation semantics.

We make the following hypothesis:

- **H4:** Speed will have the highest scores since it preserves the semantic synchronization and presents a transition process, and Hybrid will be the least preferred since it has inconsistent motions, i.e., flash.

3.2.3 Results

The statistical results of Study 2 are shown in Fig. 4. We ran normality tests on the data, and found non-normality distributions. We thus utilized Kruskal-Wallis H-test and Post-hoc Dunn's Test on the data combined from the two scenarios to examine the effects of the cases. We will elaborate on the specific results under the confidence interval of $p < 0.05$ below.

Similarity. Kruskal-Wallis H-test analysis shows a significant effect of the three conditions (H *Statistic* = 15.98, p = 3.4e-4). From the Dunn's Test, Speed (M = 5.58, SD = 0.72) and Const (M = 5.52,

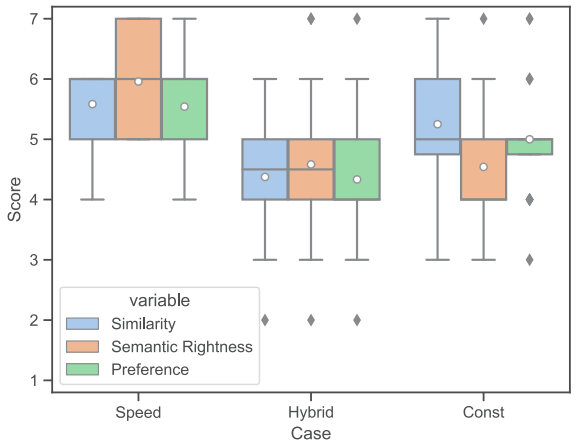

Fig. 4. The statistical results for the combined data from Scenarios 1 and 2 of Study 2.

SD = 1.11) are both significantly more "similar to the anticipation" than Hybrid (*M* = 4.38, *SD* = 1.06), with p = 8.0e-5 and p = 0.011, respectively. Speed and Const are not significantly different (p = 0.160).

Semantic Rightness. Kruskal-Wallis H-test analysis shows a significant effect of the three conditions (*H Statistic* = 22.29, p = 1.4e-5). According to Dunn's Test, Speed (*M* = 5.96, *SD* = 0.81) is significantly more right in semantics than Hybrid (*M* = 4.58, *SD* = 1.25) and Const (*M* = 4.54, *SD* = 1.02) with p = 8.7e-5 and p = 2.3e-5, respectively. But Hybrid and Const have no significant difference (p = 0.755).

Preference. Kruskal-Wallis H-test shows a significant difference among the three conditions (*H Statistic* = 16.41, p = 2.7e-4). We further ran Dunn's Test and found that Speed (*M* = 5.54, *SD* = 0.83) is significantly more preferred than Hybrid (*M* = 4.33, *SD* = 1.09) and Const (*M* = 5.00, *SD* = 0.93) with p = 5.1e-5 and p = 0.043, respectively. Const is also significantly more preferred than Hybrid (p = 0.043).

Discussion. **H4 is supported:** In the presence of delay, "Speed" is significantly better than "Hybrid" and "Const" in preserving the semantics of the interaction and is significantly more preferred. This is also confirmed by the feedback from the participants: "In Scenario 1, The TV being turned on with the avatar being still on the way makes the break in the experience very clear"; "Flash is always unacceptable". The Similarity score of "Const" remains close to "Speed" due to the most natural avatar animation. We observe that the participants perceive the delay more clearly in Scenario 1 than in Scenario 2. According to the analysis of the scenario-separated data (provided in the supplemental document), although the distribution of the scores in Scenario 1 is similar to that in Scenario 2, the effect of the conditions is not significant regarding "Similarity" in Scenario 1. This is in line with the observation in Study 1 that the delay has the primary effect on the user experience. The three conditions in Study 2 are compromises to the transition delay. The clearer perception of the delay results in the feeling that the tasks are all not that "similar to the anticipation" in Scenario 1. Besides, the Similarity score of "Const" showing a rise in Scenario 2 also indicates such a perception difference, which might lead to differences in the potential perceived delay threshold. We attribute this to the moving direction of the avatar relative to the user. This is confirmed by a comment from one participant: "I feel it not so strange when the motion is not matched with the conversation in Scenario 2 since the avatar is constantly approaching and already in the affinity when saying 'OK, show me'." To sum up, we choose the "Speed" transition style in our implementation. Note that the multiple of this speedup is determined by the ratio of the distance between the remote user and the remote target to the distance between the avatar and the local target. We implement this mechanism as illustrated in Sect. 4.3.

3.3 Study Conclusion

We examine the impact of the transition delay in Study 1, along with several other design choices, i.e., delay style, and obstacle avoidance. The results show that determining the avatar's target at the end of the remote user's transition with a large delay can do significant harm to the user's experience. Therefore, we need to predict the user's target in advance to reduce the transition delay. Moreover, once having the target, we need to generate a path for the avatar to move from the current position to the target position. This path should be safe and natural enough to lead the avatar to avoid obstacles and preserve the semantics that it is moving towards the target. From Study 2, we obtain the design choice to speed up the avatar's transition when there exists an inevitable prediction delay. It helps keep the semantics of the conversation contexts and the avatar's motion synchronized while maintaining the coherence of the transition process.

4 Method

In the previous section, we verify the necessity of the target prediction and the transition process in such AR telepresence with heterogeneous spaces, and obtain several design choices for the implementation. In this section, we will introduce our proposed Predict-and-Drive method to adapt the avatar's motion during the whole telepresence experience in real-time. Our method is comprised of two main components, i.e., the target prediction component and the avatar control component. We implement the components in a prototype system using only the HoloLens with full wireless mobility, without any additional outside-in or obtrusive setup.

4.1 Scene Understanding and Annotation

Traditionally, to construct a mapping from the objects in a space to those in another space, we first need 3D scans of the two spaces and then use automatic or semi-automatic approaches to perform semantic or even instance-level segmentation on the scanned room models [26, 32, 36]. Afterwards, we need to match them according to certain geometry or semantic factors [10, 29]. Such an understanding-matching pipeline needs too many preliminary setups and is often computationally expensive.

Instead, we adopt a less precise but fully interactive and mobile create-and-align pipeline leveraging the advantages of our AR user interface. Specifically, we create annotated 3D virtual boxes representing the main objects (e.g., sofa, TV, table) of the space, overlay them in the real world as AR contents, and let users assign and align each of them to the physical objects in the space using mid-air gestures empowered by AR HMDs. They can drag the 3D virtual boxes to the desired position and resize them to fit the edges of physical objects. The same objects are created in all the spaces involved, forming an intrinsic object-level annotation and mapping. For the physical obstacles, users can also create aligned virtual replicas but choose not to map them to each other's space. Similar manual annotation features occur recently in VR HMDs, e.g., the Presence Platform from Meta. It cannot generate an exhaustive segmentation, annotation, and mapping for the objects in a space, but is informally reported as effective enough for daily spaces since most of the installations are regularly shaped, and thus easy to align. It is reported to be entertaining for the game-like experience. We build this indexable coarse virtual replica for each space as the input and reference for the subsequent target prediction and avatar control components.

4.2 Target Prediction Component

After obtaining the basic scene representation, we first predict the target position of the avatar in the target prediction component. To explore the method for such prediction (as well as the avatar control) in our AR telepresence problem, we refer to the RDW technique in VR because we find it shares a similar goal to ours in manipulating an agent's movement with respect to two different environments in real-time. We summarize the observed connections in Table 2 and clarify them below. In RDW, the role to initiate a transition is an avatar, embodied by a user, regarding a virtual environment. The agent to control is a VR user in a physical environment. While in AR telepresence, the role to initiate a transition is a user, regarding a remote space. The agent to control is an avatar,

Table 2. Relations between VR redirected walking and AR telepresence.

	VR Redirected Walking	AR Telepresence
Goal	Manipulate an agent's movement with respect to two environments in real-time	
Role to initiate a transition	User-embodied avatar	Remote AR user
Environments	Physical-Virtual	Local-Remote
Agent to control	VR user	Avatar in AR
Coupling	Fully	Partially

representing a remote user, in a local space. RDW controllers aim to steer the VR user to avoid physical obstacles and boundaries while exploring a virtual environment, which is different from the physical one in size and configuration. The user and the embodied avatar are tightly coupled in position and point of view without any delay. From the local user's point of view, the AR telepresence avatar controller in this work aims to redirect the remote user's avatar's movement to avoid local physical obstacles and transition to the right target while the remote user is moving to a target in a dissimilar remote space. The avatar and the remote user should be semantically coupled with little delay as well. With these correspondences, we can see the AR telepresence problem as an RDW task with more semantic constraints, and derive potential solutions from the existing RDW techniques.

As existing approaches in predictive RDW controllers still cannot handle AR scenes or need history trajectories [3, 27, 41], we adopt a reactive moving-direction-based target prediction method similar to those introduced in RDW controllers to provide haptic feedback [27]. In order to enable low-error prediction with no need for prerequisite user paths, we design a method based on the spatial-temporal relation between the current user's moving direction and the visibility polygon [28] of the space. This component takes the remote user's position and the spatial index as the input. The user's position is tracked in real-time through the HoloLens onboard head tracking system, and the position and rotation of each object are obtained from the pre-constructed virtual replica of the user's space.

The visibility polygon (VP) is a representation of the 2D area that is visible to the user, changing accordingly with the user's position. We assume that when the user is moving to an object, the object will be visible to the user sooner or later during the transition. Therefore, the VP would be where the potential target lies. Specifically, the VP consists of a sequence of vertices ordered counterclockwise, between which edges are connected sequentially. The notion of a VP slice is defined to be a triangular area formed by the user's position and two other consecutive vertices if they are not co-linear. Here we define an object slice as the triangle area formed by the user's position and the first and the last vertices that belong to the object, as illustrated in Fig. 5. Since our objects are all represented as boxes, thus rectangles in 2D, by definition, an object slice can be a single VP slice that covers the object or a combination of two VP slices, depending on whether the first and the last vertices are on the same edge or diagonal. Specifically, to predict the target object, we construct and update the object slices in real-time. We define an object as the target if the user's moving direction stays in the object's slice for a certain period of time (0.5s in our implementation).

When the user moves straight to the target, the basic prediction method described above can function well. However, for the situation where there exists an obstacle between the user's starting point and the target point, it will have a relatively big delay since the target can only be predicted at the latter part of the transition, as illustrated in Fig. 6. This is inevitable since the former part of the transition is ambiguous. To reduce the prediction delay caused by such ambiguity, we further design a grouping component as an upgrade to the basic target prediction component. Instantly at the beginning of the transition, it finds the group of objects the remote user is approaching, denoted as $Group\{Obj_1, Obj_2 ... Obj_n\}$, i.e., the distance D_{RU-i} between the remote user and Obj_i reduced compared to that in the last frame. Then, we use the normalized distance reduction as the weight to calculate a weighted center of the group as a coarse remote virtual target, as

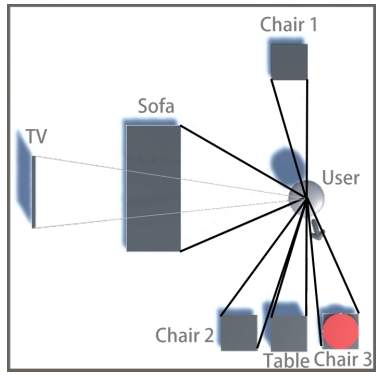

Fig. 5. An example for the object slice visualized in a space from the top-down view. The object slices with bold black edges are considered visible to the user while the thin grey ones are considered invisible.

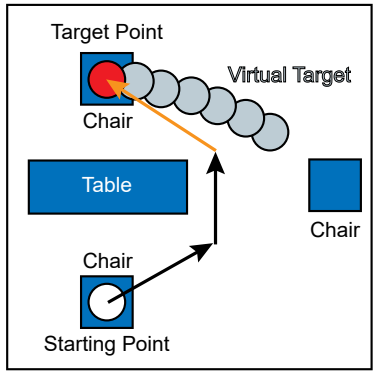

Fig. 6. An example to explain the situation when grouping is needed. The semantic ambiguity of the locomotion in the former part (black arrows) makes it hard to determine the target at this stage (e.g., the target could be the chair on the right or at the top). The target is cleared only in the latter part of the transition (the orange arrow). With the grouping component, the controller can update the virtual target every frame with zero delay, which will consistently approach (during the black arrows) and eventually reach (during the orange arrow) to the real target.

illustrated in Equation 1. $PObj_i$ and $PCenter$ represent the positions of the i-th object in $Group\{Obj_1, Obj_2 ... Obj_n\}$ and the virtual target, respectively. We have

$$PCenter = \sum_{j=1}^{n} PObj_j * \frac{D_{RU-j}^{last\,frame} - D_{RU-j}^{this\,frame}}{\sum_{i=1}^{n}(D_{RU-i}^{last\,frame} - D_{RU-i}^{this\,frame})}. \quad (1)$$

Since the objects are pre-matched between the two spaces, the detected remote target group also has a corresponding local object group. Assigning the same weights in the remote group to the corresponding local objects, we calculate the weighted center of the local group as the local virtual target for the avatar. It enables the avatar to also start moving without any delay when the remote user initiates the transition. The virtual target is calculated and updated in every frame in real-time.

4.3 Avatar Control Component

When we get the remote user's target position, we can drive the avatar to the corresponding local target. The purpose of the avatar control component is to generate a natural and semantically right path for the avatar, leading it to the local target simultaneously with the remote user while avoiding physical obstacles in the local space. It takes the spatial index and the local target position as the input. The spatial index storing the position, orientation, and size of each object is obtained from the pre-constructed virtual replica (Sect. 4.1) of the local space, and the local target position is generated by the target prediction component.

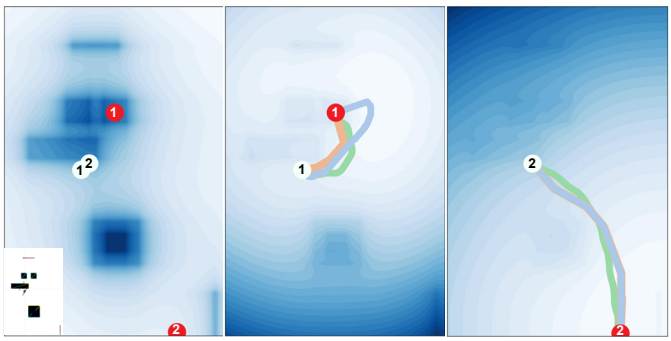

Fig. 7. The potential field of the experiment space in Fig. 2. The thumbnail in the bottom-left corner of (a) shows the top-down view of the space with the avatar in it. (a) shows the APF of the whole space with repulsive forces only. Circles with numbers denote the starting point (white) and the target point (red) for Scenarios 1 and 2. (b) and (c) show the ground-truth path (green), the path generated by our basic method (orange), and the path generated by our grouping method (blue) for Scenarios 1 and 2, respectively. In (c), the Basic and Grouping paths overlay with each other since there are fewer objects along the way to the target object than the situation in Scenario 1.

We use the Artificial Potential Field (APF) method to control the avatar's moving direction. It is a well-studied method in the robotics motion planning problem, where the robot needs to plan its path from the starting point to a given target point while avoiding collisions with obstacles [12]. Researchers in the field of RDW also adopt the APF algorithm as the steering controller [30]. The idea of the APF algorithm is to construct an energy field in which the energy E_p of each position p is calculated by the potential function. The potential function consists of two parts, namely the repulsive forces from the obstacles and the attractive force from the target. In our case, since the obstacles in the space are all represented by boxes, we calculate the repulsive forces of each obstacle with respect to all its four edges [9].

We construct and update the APF of the local space (the AR Scene shown in Fig. 2) following the formulations provided in the supplemental file, and visualize it in Fig. 7. It determines the avatar's moving direction in the next frame according to the negative gradient of the field at the current position of the avatar. The potential field changes in real-time with the variation of the target position, which is determined by the target prediction component, constantly driving the avatar until reaching the target. Note that Fig. 7 (b) and (c) show the final APF after the prediction module settles the target, before which there is a delay in the Basic method and the virtual target is constantly changing in the Grouping method.

For the avatar's moving speed, according to the results in our Pilot Study 2, we dynamically adjust both its translation rate and animation playback rate to ensure that the avatar and the remote user reach the respective targets at the same time. Let $D_{Avatar-Target}$ be the distance between the avatar and its local target, and $D_{User-Target}$ be the distance between the remote user and the remote target. We multiple the avatar's moving speed and playback rate both by $D_{User-Target}/D_{Avatar-Target}$ and update them every frame.

5 Initial Evaluation

We design the evaluation experiment with an ablation of the grouping component and a comparison with the ground truth (i.e., the real-user-recorded path in the local space, representing the most natural avatar path adapted from the remote user's transition). In the accompanying video, we also demonstrate the grouping component's ability to further reduce the transition delay by visually comparing the avatar transitions generated by our basic method and the upgraded method with dynamic grouping in three simulated space pairs. The evaluations verify the effectiveness of our proposed Predict-and-Drive method, the performance improvement of the upgraded method with dynamic grouping, and the overall closeness to the ground truth.

5.1 Initial User Study

5.1.1 Participants

We recruited 12 undergraduate students (8 males, 4 females; average age: 23.3 (SD = 0.97)) from the local campus as the participants. Most of them had a moderate level of AR/VR experience: 11 had experienced several times, and 1 is an AR/VR developer.

5.1.2 Experiment Setup

We record real-user trajectories for both scenarios of the user in the remote space. Then we load each of them as the remote user's transition while activating our target prediction and avatar control components to adapt the avatar's motion. We record the adapted motions with and without the dynamic grouping component, and present them along with the ground-truth paths of the avatar recorded from real human in Study 1 and Study 2. All the three transitions take 8s since we limit the remote user's real transition to 8s, and the avatar's speed is dynamically adapted (as illustrated in Sect. 4.3).

5.1.3 Study Cases

As mentioned above, we consider three conditions, the ground truth **GT**, the motion sequence generated without dynamic grouping **Basic**, and the motion sequence with dynamic grouping **Grouping** in this study. The paths for the 3 experiment conditions are shown in Fig. 7. As shown in Fig. 7 (b), the Basic method generates a path more similar to the ground truth than that generated by Grouping, which, takes a turn around the target. However, the Basic path has a start delay (around 5s in Scenario 1 and 3s in Scenario 2) because with the existence of the obstacle the basic method needs some time to determine the target while the Grouping path has zero delay. The visual effects of the three conditions are shown in the supplementary video.

We make the following hypotheses:

- **H5**: Both the basic and grouping cases will have significantly higher scores regarding all the three metrics compared to the previous method, i.e., the "Flash-WIP" case in Pilot Study 1.
- **H6**: Grouping will further improve the user's experience to a level closer to GT than Basic.

5.1.4 Results

The statistical results of the experiment are summarized in Fig. 8. We ran normality tests on the data, and found non-normality distributions. We thus conducted Kruskal-Wallis H-test and Post-hoc Dunn's Test on the data combined from the two scenarios to check the significance of the effects of the cases. All of the results below are reported under the confidence interval of $p < 0.05$.

Similarity. Kruskal-Wallis H-test shows a significant difference among the three conditions ($H = 33.45$, $p = 5.5e\text{-}8$). We further ran a Dunn's Test and found that GT ($M = 6.29$, $SD = 0.75$) is significantly more "similar to the anticipation" than Basic ($M = 4.33$, $SD = 0.92$) with $p = 6.5e\text{-}8$, Grouping is significantly more "similar to the anticipation" than Basic ($p = 7.2e\text{-}6$), while GT and Grouping ($M = 5.86$, $SD = 1.08$) have no significant difference ($p = 0.359$).

Semantic Rightness. Kruskal-Wallis H-test shows a significant effect of the three conditions ($H = 27.27$, $p = 1.2e\text{-}6$). According to Dunn's Test, GT ($M = 6.29$, $SD = 0.69$) is significantly more right in semantics than Basic ($M = 4.54$, $SD = 1.06$) with $p = 8.6e\text{-}7$, and Grouping ($M = 5.96$, $SD = 1.08$) and Basic have a significant difference ($p = 7.1e\text{-}5$). But GT and Grouping have no significant difference ($p = 0.343$).

Preference. Kruskal-Wallis H-test shows a significant effect of the three conditions ($H = 31.52$, $p = 1.4e\text{-}7$). GT ($M = 6.17$, $SD = 0.87$) is significantly more preferred than Basic ($M = 4.25$, $SD = 0.74$) with $p = 1.1e\text{-}7$. Grouping ($M = 5.75$, $SD = 1.22$) is significantly more preferred than Basic ($p = 2.5e\text{-}5$). Grouping and GT have no significant difference ($p = 0.272$).

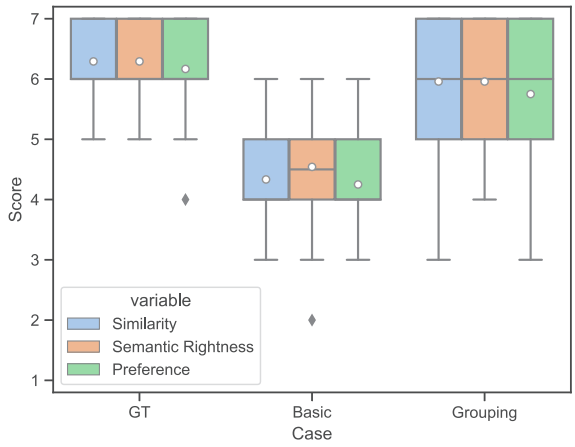

Fig. 8. The statistical results for the combined data from Scenarios 1 and 2 of the user study in the Initial Evaluation.

Discussion. **H5 and H6 are both supported:** "Grouping" provides a decent experience with no significant difference to "GT" regarding all the metrics in both Scenarios 1 and 2. "Basic" gets higher mean scores than "Flash-WIP" in Pilot Study 1, though we cannot verify the significance level due to the difference in the participant composition and experiment cases. The grouping component can significantly improve the performance of the basic method. We observe that in Scenario 1, where Grouping's trajectory deviates more from GT's, some participants perceive such a deviation as "taking a turn to better sit on the chair," while others feel "it is going somewhere else before turning back to the chair". It indicates the possible existence of a detection threshold for such deviations to be imperceptible or reasonable. The comparison with the avatar transition effect of the previous method [40] is also shown in Fig. 7 (b) and (c). The previous method generates transitions with the maximum delay (8s) and a flash from the starting point and the target point, while ours can reduce ("Basic") and further nearly eliminate ("Grouping") the delay and present a complete transition process with the user-preferred style. In the supplemental materials, we also provide the analysis of the scenario-separated data and the visual comparison between the "Flash-WIP" case in Pilot Study 1 (demonstrating the effect of the previous method) and the "Grouping" case.

5.2 Visual Comparison in Simulation Environments

To further demonstrate the advantage of the proposed grouping component, we construct three simulated local-remote space pairs in the Unity simulation environment. Similar to the setup in the user study, we record a series of the remote user's movement (controlled by the keyboard) and load them while activating our target prediction and avatar control components to adapt the avatar's motion. We provide three views (i.e., the static side view, the bird view, and the avatar following view) to facilitate easy comparisons of the avatar's motions with and without the grouping component. This additional experiment shows that our grouping component can significantly reduce the delay to nearly zero in almost all the cases. The video is provided in the supplementary material.

6 Conclusion, Limitations, and Future Work

In this paper, we have explored avatar motion adaption in room-scale AR telepresence with heterogeneous spaces. We have conducted two pilot studies to figure out the problem of avatar transition delay, explore the impact of such delay, and identify the most preferred transition style for natural interactions. To help generate smooth avatar motions, we have introduced and applied a VR RDW concept to our AR telepresence scenario. Based on the findings and concept, we have proposed the Predict-and-Drive avatar motion adaption controller in the prototype system, which consists of a user target prediction component and an avatar transition control component. To help further eliminate the transition delay, we have designed a grouping component as an upgrade to the basic target prediction component. With this controller, the user's target can be predicted for the avatar's dynamic motion planning, the avatar transition delay can be diminished to a large degree, and the avatar transition can be generated smoothly in real-time. The effectiveness and performance of the controller have been validated by an evaluation experiment. Next we address the limitations of our work and the corresponding potential future work.

Generalization. Our current conclusions are based on relatively small numbers of participants with limited background diversity. We are interested in conducting larger-scale studies involving more participants with greater background diversity in the future. For a better generalization, we can explore more factors (e.g., the type and size of the spaces) that could influence the design choice of the telepresence avatar's motion and include more remote-local space pairs in the evaluation. Besides, estimating the common transition time in room-scale daily AR telepresence and giving an according taxonomy of activities in such scenarios could benefit the generalization and future exploration.

Target Prediction. Our current target prediction component yields a particular object as the target. We can explore new methods for predicting the remote user's future trajectory, which might be more generalized since sometimes the remote user can target at a position instead of an object. We can leverage the gaze direction to enhance the robustness of our grouping component for the situation where random slight displacements occur when the user is not really moving [27]. We can also explore more advanced methods such as fully utilizing the information in the visibility polygon, using reinforcement-learning-based approaches, and considering the user's habitual preferences and relation to surrounding objects as in the study of multi-player networked games [20]. Besides, the threshold for the moving direction staying in an object slice can be further examined for various scene installations.

Short-term Motions. We focus on the scenarios with a relatively long-term transition between two positions. For short motions with no transition issue (like getting up or sitting down), we can incorporate pose estimation (based on vision or HMD IMU data classification) to change the avatar's pose with regard to the remote user's pose in real-time. For motions with more delicate short-distance transitions, if they are semantically changing the target position (e.g., moving from one to the other side of a whiteboard), a more detailed and effective scene understanding [35, 37] can allow our method to work still.

Multiple Users and Spaces. In this paper, we mainly consider the scenario consisting of two users separated in two heterogeneous physical spaces. We believe it is a fundamental and common scenario in AR telepresence. For scenarios involving multiple users in two or multiple spaces, semantically mapping the objects among multiple spaces could be a challenge. We could consider augmenting or diminishing some objects in each space to achieve the mapping. The system should also respect some social interaction manners, such as enabling private group chats in multi-user experience leveraging spatial audio or system logic level implementation, instead of broadcasting all the conversations. Besides, the real-time performance of the controller needs further testing with the increasing numbers of users and spaces.

Acknowledgments

We thank the anonymous reviewers for their constructive comments, and the interview and user study participants for their time. This work was supported by grants from National Key Research and Development Program of China (No. 2020AAA0108800), the City University of Hong Kong (No. 7005729, 9667234, 7005590, 9229094), the National Natural Science Foundation of China (No. 62172326 and 62137002), Innovative Research Group of the National Natural Science Foundation of China (No. 61721002), Innovation Research Team of Ministry of Education (No. IRT 17R86), Project of China Knowledge Centre for Engineering Science and Technology, Project of Chinese academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China". Hui Ye was supported by the RGC Postdoc Fellowship Scheme.

REFERENCES

[1] B. Benda and E. D. Ragan. The effects of virtual avatar visibility on pointing interpretation by observers in 3d environments. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 50–59. IEEE, 2021.

[2] K. Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 981–992, 2016.

[3] Z.-Y. Chen, Y.-J. Li, M. Wang, F. Steinicke, and Q. Zhao. A reinforcement learning approach to redirected walking with passive haptic feedback. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 184–192. IEEE, 2021.

[4] Y. Choi, J. Lee, and S.-H. Lee. Effects of locomotion style and body visibility of a telepresence avatar. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1–9. IEEE, 2020.

[5] M. Gonzalez-Franco, B. Cohn, E. Ofek, D. Burin, and A. Maselli. The self-avatar follower effect in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 18–25. IEEE, 2020.

[6] M. Gottsacker, N. Norouzi, K. Kim, G. Bruder, and G. Welch. Diegetic representations for seamless cross-reality interruptions. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 310–319. IEEE, 2021.

[7] S. Hirche and M. Buss. Passive position controlled telepresence systems with time delay. In *Proceedings of the 2003 American Control Conference, 2003.*, vol. 1, pp. 168–173. IEEE, 2003.

[8] Y. Huang and M. Kallmann. Planning motions and placements for virtual demonstrators. *IEEE transactions on visualization and computer graphics*, 22(5):1568–1579, 2015.

[9] Y. K. Hwang, N. Ahuja, et al. A potential field approach to path planning. *IEEE Transactions on Robotics and Automation*, 8(1):23–32, 1992.

[10] D. Jo, K.-H. Kim, and G. J. Kim. Spacetime: adaptive control of the teleported avatar for improved ar tele-conference experience. *Computer Animation and Virtual Worlds*, 26(3-4):259–269, 2015.

[11] M. Keshavarzi, A. Y. Yang, W. Ko, and L. Caldas. Optimization and manipulation of contextual mutual spaces for multi-user virtual and augmented reality interaction. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 353–362. IEEE, 2020.

[12] O. Khatib. Real-time obstacle avoidance for manipulators and mobile robots. In *Autonomous robot vehicles*, pp. 396–404. Springer, 1986.

[13] D. Kim, J.-e. Shin, J. Lee, and W. Woo. Adjusting relative translation gains according to space size in redirected walking for mixed reality mutual space generation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 653–660. IEEE, 2021.

[14] K. Kim, D. Maloney, G. Bruder, J. N. Bailenson, and G. F. Welch. The effects of virtual human's spatial and behavioral coherence with physical objects on social presence in ar. *Computer Animation and Virtual Worlds*, 28(3-4):e1771, 2017.

[15] Y. Kim, H. Park, S. Bang, and S.-H. Lee. Retargeting human-object interaction to virtual avatars. *IEEE transactions on visualization and computer graphics*, 22(11):2405–2412, 2016.

[16] Y. Lang, W. Liang, and L.-F. Yu. Virtual agent positioning driven by scene semantics in mixed reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 767–775. IEEE, 2019.

[17] M. E. Latoschik, D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch. The effect of avatar realism in immersive social virtual realities. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2017.

[18] J. Lee, M. Lee, G. J. Kim, and J.-I. Hwang. Effects of virtual gait visualization in walk-in-place on body ownership and presence. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2020.

[19] N. H. Lehment, D. Merget, and G. Rigoll. Creating automatically aligned consensus realities for ar videoconferencing. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 201–206. IEEE, 2014.

[20] S. Li, C. Chen, and L. Li. A new method for path prediction in network games. *Computers in Entertainment (CIE)*, 5(4):1–12, 2008.

[21] A. Maimone, X. Yang, N. Dierk, A. State, M. Dou, and H. Fuchs. General-purpose telepresence with head-worn optical see-through displays and projector-based lighting. In *2013 IEEE Virtual Reality (VR)*, pp. 23–26. IEEE, 2013.

[22] N. McHenry, J. Spencer, P. Zhong, J. Cox, M. Amiscaray, K. Wong, and G. Chamitoff. Predictive xr telepresence for robotic operations in space. In *2021 IEEE Aerospace Conference (50100)*, pp. 1–10. IEEE, 2021.

[23] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, p. 741–754. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2984511.2984517

[24] T. Pejsa, J. Kantor, H. Benko, E. Ofek, and A. Wilson. Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pp. 1716–1725, 2016.

[25] G. Samaraweera, R. Guo, and J. Quarles. Latency and avatars in virtual environments and the effects on gait for persons with mobility impairments. In *2013 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 23–30. IEEE, 2013.

[26] T. Shao, W. Xu, K. Zhou, J. Wang, D. Li, and B. Guo. An interactive approach to semantic modeling of indoor scenes with an rgbd camera. *ACM Transactions on Graphics (TOG)*, 31(6):1–11, 2012.

[27] F. Steinicke, G. Bruder, L. Kohli, J. Jerald, and K. Hinrichs. Taxonomy and implementation of redirection techniques for ubiquitous passive haptic feedback. In *2008 International Conference on Cyberworlds*, pp. 217–223. IEEE, 2008.

[28] S. Suri and J. O'Rourke. Worst-case optimal algorithms for constructing visibility polygons with holes. In *Proceedings of the second annual symposium on Computational geometry*, pp. 14–23, 1986.

[29] T. Tahara, T. Seno, G. Narita, and T. Ishikawa. Retargetable ar: Context-aware augmented reality in indoor scenes based on 3d scene graph. In *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 249–255. IEEE, 2020.

[30] J. Thomas and E. S. Rosenberg. A general reactive algorithm for redirected walking using artificial potential functions. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 56–62. IEEE, 2019.

[31] M. Usoh, K. Arthur, M. C. Whitton, R. Bastos, A. Steed, M. Slater, and F. P. Brooks Jr. Walking> walking-in-place> flying, in virtual environments. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 359–364, 1999.

[32] J. Valentin, V. Vineet, M.-M. Cheng, D. Kim, J. Shotton, P. Kohli, M. Nießner, A. Criminisi, S. Izadi, and P. Torr. Semanticpaint: Interactive 3d labeling and learning at your fingertips. *ACM Transactions on Graphics (TOG)*, 34(5):1–17, 2015.

[33] M. E. Walker, D. Szafir, and I. Rae. The influence of size in augmented reality telepresence avatars. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 538–546. IEEE, 2019.

[34] N. L. Williams, A. Bera, and D. Manocha. Redirected walking in static and dynamic scenes using visibility polygons. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4267–4277, 2021.

[35] Q. Xie, Y.-K. Lai, J. Wu, Z. Wang, Y. Zhang, K. Xu, and J. Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10447–10456, 2020.

[36] K. Xu, H. Huang, Y. Shi, H. Li, P. Long, J. Caichen, W. Sun, and B. Chen. Autoscanning for coupled scene reconstruction and proactive object analysis. *ACM Transactions on Graphics (TOG)*, 34(6):1–14, 2015.

[37] T. Yin, X. Zhou, and P. Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.

[38] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, and W. Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 547–556. IEEE, 2019.

[39] L. Yoon, D. Yang, C. Chung, and S.-H. Lee. A full body avatar-based telepresence system for dissimilar spaces. *arXiv preprint arXiv:2103.04380*, 2021.

[40] L. Yoon, D. Yang, J. Kim, C. Chung, and S.-H. Lee. Placement retargeting of virtual avatars to dissimilar indoor environments. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[41] M. Zank and A. Kunz. Using locomotion models for estimating walking targets in immersive virtual environments. In *2015 International Conference on Cyberworlds (CW)*, pp. 229–236. IEEE, 2015.