Understanding the Quality of User Experience in Telepresence Systems From an Information Theory Perspective

Ruiqing Wang, *Student Member, IEEE*, Kebin Liu, *Senior Member, IEEE*, Ziyue Dang, *Student Member, IEEE*, Xu Wang, *Member, IEEE*, Fan Dang, *Member, IEEE*, Yue Sun, *Student Member, IEEE*, Yuang Tong, *Student Member, IEEE*, Haitian Zhao, *Member, IEEE*, and Yunhao Liu, *Fellow, IEEE*

***Abstract*—Efforts to enhance the user experience (UX) of telepresence edge systems in various application scenarios have been significant. However, existing approaches tend to focus on specific aspects, leaving us with a fragmented understanding of UX quality. We address this gap by examining Video Conference Systems (VCSs) for remote collaboration, using an Information Theory Perspective as a lens. We introduce a novel model to quantify the multimodality information users receive while engaged in mobile office environments, enabling an evaluation of existing VCSs. Our approach transforms the assessment of UX quality into the measurement of a set of information *channels*. Based on this insight, we identify new prospects and meaningful guidelines for future multimedia telepresence edge systems, and try to induce a new prototype design under cost restriction. To demonstrate the validity of our method, we implement the prototype which seamlessly integrates visual, audio, and olfactory dimension information. Extensive experiments and user studies validate the effectiveness and practicality of our approach.**

***Index Terms*—Telepresence edge system, model analysis, user experience quality.**

I. Introduction

TELEPRESENCE edge systems have been increasingly common in people's daily lives. More and more conversations, work meetings as well as formal conferences are moving from offline to online. Significant attention has been drawn from both academia and industry to improve the efficiency and user experience (UX) of these systems, especially video conference systems (VCSs) for remote collaboration. As shown in Fig. 1 (a), cloud-based approaches such as Zoom [1] and Tencent Meeting [2] tried to provide high-resolution images and voice streams with low latency. MAJIC [3], Hydra [4], Gaze-2 [5] and Multi-View [6], [7] discussed the importance of conveying gaze direction as illustrated in Fig. 1 (b). The fixed viewpoint can result in a poor sense of reality, thus many efforts [8], [9], [10], [11], [12] have been devoted to capturing and displaying 3D user silhouettes. Fig. 1 (c) demonstrates a stereoscopic solution using a cylindrical display. As shown in Fig. 1 (e), Horizon Workrooms [13] from Meta enabled people to collaborate in virtual reality with their cartoon avatar, and robot surrogates [14], [15] such as the one shown in Fig. 1 (d) was also introduced to appear in remote conference rooms. The What-You-See-Is-What-I-See [16], [17], [18] VCSs focused on building a shared "screen" for both co-located and remote users, where users' images were embedded into physical and digital contexts for deictic referencing. An example is demonstrated in Fig. 1 (f).

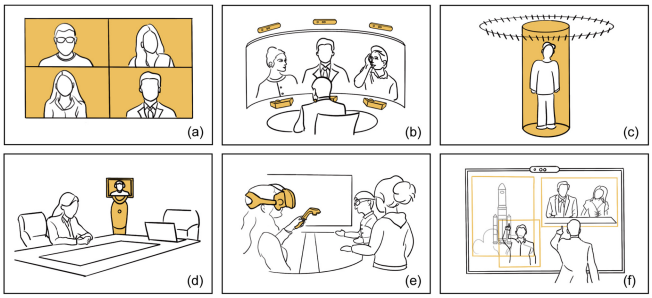


Fig. 1. Visualization of typical telepresence edge systems: (a) Cloud-based approach (b) Mutual gaze aware solution (c) Holographic display for 3D reconstruction (d) Robotic avatar (e) VR collaboration (f) WYSIWIS meeting.

After reviewing all these mainstream techniques, we find that each of them improves the telepresence edge system from a particular perspective and puts emphasis on limiting factors. We, however, still lack a global picture and UX quality measurement method of them. A novel descriptive model is needed for systematic measurement and analysis.

We take the user experience in local meetings, in other words, "being there", as a reference for a "good" mobile officing system. The insights behind this choice are that human beings perceive their surroundings through all of their sensory organs. For example, in an offline meeting, people

Manuscript received 16 July 2023; revised 2 November 2023; accepted 3 December 2023. Date of publication 10 January 2024; date of current version 14 August 2025. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0114903, and in part by the National Natural Science Foundation of China (NSFC) under Grant 62202263. *(Corresponding author: Kebin Liu.)*

Ruiqing Wang, Kebin Liu, Xu Wang, Fan Dang, Yue Sun, Yuang Tong, and Yunhao Liu are with the Global Innovation Exchange, Tsinghua University, Beijing 100084, China (e-mail: wang-rq22@mails.tsinghua.edu.cn; kebinliu2021@tsinghua.edu.cn; xu_wang@tsinghua.edu.cn; dangfan@tsinghua.edu.cn; suny21@mails.tsinghua.edu.cn; tya21@mails.tsinghua.edu.cn; yunhao@tsinghua.edu.cn).

Ziyue Dang is with the Computer Science Department, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: ziyue.dang@cs.ucla.edu).

Haitian Zhao is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: zhaoht2022@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TCE.2024.3352240

1558-4127 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

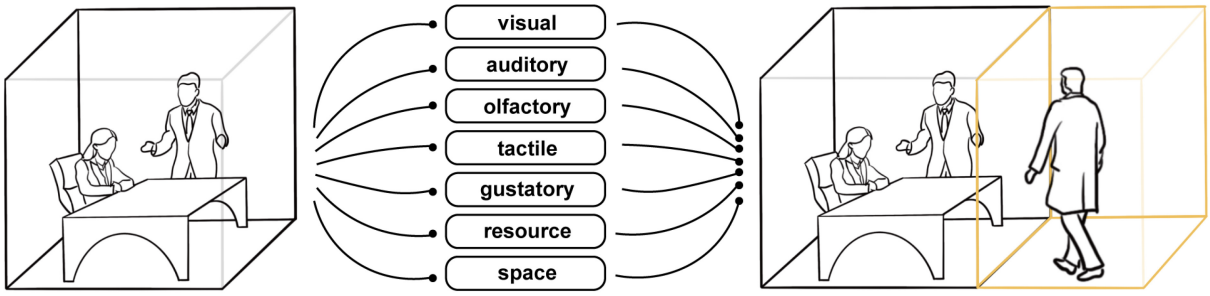


Fig. 2. Observation of this study: the telepresence edge system can be regarded as a set of information *channels* connecting distributed users.

see each other with their eyes and hear other participants' voices with their ears. Each sensory organ corresponds to a particular modality of information. Therefore, the dimensions of input information to the human brain are limited by the number of sensory organs. Information from all sensory organs forms a complete set of inputs to the human brain, which thus determines the user's experiences. Based on this observation, we propose to revisit VCSs, the most typical telepresence edge system from an Information Theory [19] perspective. It is worth mentioning that our model is different from the traditional Information Theory which calculates the channel capacity of computer networks. We hope to model the information received by all sensory organs of the user during the remote collaboration, which means the information flow is from user to machine to user. As illustrated in Fig. 2, in an online meeting scenario, the separation of physical space blocks people from perceiving information with their sensory organs directly from remote sites. In this case, VCS takes the responsibility of sampling, encoding, transmitting, and finally demonstrating multi-modality information from one meeting room to another. It can be regarded as a set of *channels* for different sensory information.

Besides, as finding the trend that digital media can offer richer information and achieve higher efficiency than just "being there," we add new *channels* to our model to take these features into consideration. The concept of "beyond being there" has been proposed in [20], and recent advances in computation speed, AI algorithm functionality, *etc.*, bring us new possibilities such as verbal content visualization [21], auto-summary [22], Air-Text [23] and virtual raise hand (VRH) prediction [24] to mobile officing applications. In all, with this model, the problem of studying the efficiency and UX of telepresence edge systems can be transformed into measuring the quality of each *channel*.

The contribution of this study can be summarized as follows.

- First, we propose a novel Information Theory based model to calculate the multi-modality information received by users and to evaluate the UX quality of telepresence edge systems.
- Second, we analyze 29 typical solutions and discuss insights with heuristic guidance for further remote collaboration improvements.
- Third, we propose a design under the above guidelines that aim to achieve high UX quality at a modest cost.
- Fourth, we implement the prototype and conduct extensive experiments to validate the validity of our methods.

The rest of this paper is organized as follows. Section II describes research that is related to our work. We present an Information Theory based model for UX quality measurement in Section III. The insights from our analysis as well as meaningful and heuristic guidelines for future design are discussed in Section IV, and we also present a prototype design in this section. We express the prototype system implementation and experimental results in Section V, then conclude our work and its future plans in Section VI.

II. Related Work

A telepresence edge system combines high-definition video, audio, and interactive components to create a unique "face-to-face" experience on the Web. The most typical telepresence edge system is VCSs for mobile officing.

VCSs that directly capture, transmit and present images at a specific and fixed angle include MAJIC [3], Hydra [4], GAZE-2 [5] and Teleport [25]. All these methods try to convey multiple eye contact and support proper awareness of gaze direction among the participants, and each person is represented by a separate camera/projector or camera/monitor pair. These approaches are limited by the number of devices and the fixed viewing perspective, making it difficult to expand to multiple parties.

To better convey non-verbal cues, researchers capture and reconstruct 3D scenes and 360° videos [26], [27] in immersive multimedia systems. 3D reconstruction is carried out through multiple 2D views by an area-based stereo or block matching algorithm, disparity estimation, rectification and hybrid recursive matching strategy. SphereAvatar [28] and Spherical display [29] use a spherical display for projection to achieve a 360-degree horizontally visible, perspective-correct, and life-size image. Coliseum [30] uses head tracking and IBVH for real-time rendering. Telehuman1 [10] and Blue-C [12] use polarized projection so that the user wears shutter glasses to obtain stereoscopic perception. But the above methods are essential "one-to-one" or "one-to-many" modes. In addition, C1x6 [31] employs six customized DLP projectors for fast time-sequential image display in combination with polarization. The Office of The Future [32] projects high-resolution graphics and text to any regular or irregular visible surface, such as walls, furniture, or people. Recently, VirtualCube [8]

uses six RGBD cameras to capture multi-view stereo for more accurate depth estimation and then renders high-quality videos on a surrounding life-size display by using Lumi-Net. Two or three remote users can correctly preserve mutual eye gaze, sense each other's attention, have side discussions, and also share their "workspace".

Autostereoscopic displays applied in telepresence edge systems can be divided into two categories: flat panel displays (FPDs) and curved surface displays (CSDs). FPDs include retroreflective displays, parallax barrier displays, and lenticular lens displays. MultiView [6], [7], Varrier [33] and Multi-View Lenticular Display [34] use the above three technologies, respectively, so that multiple users can view the same stereoscopic scene from different perspectives without wearing any devices, but the overall brightness of the images is low. Curved surface displays offer the opportunity to achieve continuous horizontal motion parallax and holographic scenes. In the cylindrical VCS [35], nine cameras on the remote end correspond to nine projectors on the local end. One-to-Many 3D VCS [36] projected 3D face onto any curved surface, and the visible field of view is more than 180 degrees. Telehuman2 [11] and Lightbee [37] set up a projector array around a life-size cylindrical light field display. The angle between each projector is 1.3 degrees, which is less than the average distance between adult pupils.However, such designs are costly, difficult to deploy, and currently only support "many-to-one" mode. The sense of ritual also lacks.

2D spatial sharing and fusion is a research path in VCS that follows the principle of "what you see is what I see" (WYSIWIS). From ClearBoard's [18] two separate parties sharing the same drawing board, to HyperMirror's [17] multiple images mirroring and fusion, to MirrorBlender's [16] arbitrarily adjustment of cloud conference interface. The layout, position, scale, and transparency of multi-interfaces can be adjusted by each user. This idea could also be combined with physical products, such as the notion of proxemic transitions [38], which have been proposed to guide shape-changing furniture design for informal workplace meetings.

Besides, two main types of telepresence edge robots are mobile robots and humanoid robots. Mobile robots generally have a display screen placed in the center part and can be controlled by the remote user while communicating with the local end. The MRP system used in Mobile Remote Presence [39] in the Workplace is a Texai Alpha [40] prototype. A remote user controls it by using a Web browser. Humanoid robots such as Animatronic Shader Lamps Avatars [41] and Geminoid HI-1 [42] focus on correctly reproducing "non-verbal cues" presented by the remote user's head, in order to enhance their sense of presence and reality to the local users. However, an "uncanny valley effect" may easily happen to these robots. The ThirdEye [43] focus on gaze cues transmits and creates an eye-shaped add-on display that represents a remote participant's gaze direction. This is a simple but effective way to lead a local observer's attention toward objects in the surrounding environment.

With the rapid development of XR, it has become an emerging way for users to enter and interact freely with others in the virtual world as avatars, such as in Horizon Workrooms [13] and Mesh for Teams. However, long-time use of HMDs very easy to causes fatigue, or even motion sickness, because of the desynchrony of the human brain's motion instructions and sensory feedback. More importantly, it is difficult to transmit users' facial expressions, "non-verbal cues" and the state of perception when they are wearing HMDs, which will greatly reduce the authenticity of the VCSs. Although Holoportation [9] can solve this problem, it requires a large number of specialized and expensive hardware devices. Available scenarios are limited due to the huge amount of data communication and transmission.

In this work, we hope to draw a whole picture of the telepresence edge systems for remote collaboration, evaluate their UX quality systematically, and propose a more preferable cross-modality information fusion solution.

III. An Information Theory Model for UX Quality Measurement

A. Model Definition

In our model, a telepresence edge system is defined as a vector:

$$\vec{T} = \left(C_v, C_a, C_o, C_t, C_g, C_r, C_s\right). \tag{1}$$

Each element in $\vec{T}$ refers to a *channel*, corresponding to the information dimension perceived by a particular human sensory organ. In other words, each *channel* can be regarded as a subsystem of each telepresence edge approach that is in charge of conveying a certain modality of information. We consider different information dimensions separately and assign a *channel* to each of them. In detail, C_v denotes the subsystem which distributes visual information to remote participants; C_a is the *channel* for auditory information; C_o denotes the olfactory *channel*; C_t corresponds to the tactile sensation; and, C_g is the gustatory *channel*. We can find that the first five variables in $\vec{T}$ have covered all the information types processed by human sensory organs. As we mentioned in Section I, our study also considers richer information [44] beyond traditional local meetings, provided by advanced VCS. In this study, we consider and include another two dimensions in our model, that is, the shared resources *channel* C_r and the awareness of space *channel* C_s. In our model, the shared resources include three kinds of information. First, the digital materials for demonstration, such as a picture or a 3D model. Second, the marks and deictic gestures made by the user are usually blended with digital materials in remote collaboration to convey more information such as emphasis and scope. Third, the information generated by AI algorithms, such as an auto-summarization of a user's statement and recognition of the atmosphere of the venue. The C_s *channel* indicates the awareness of space, in other words, the capability of making distributed users feel that they are located in the same physical space.

B. Channel Analysis

We evaluate the quality of each *channel* with two metrics: the *capacity* and the *cost*. The concept *capacity* comes from Information Theory, which denotes the maximum amount of

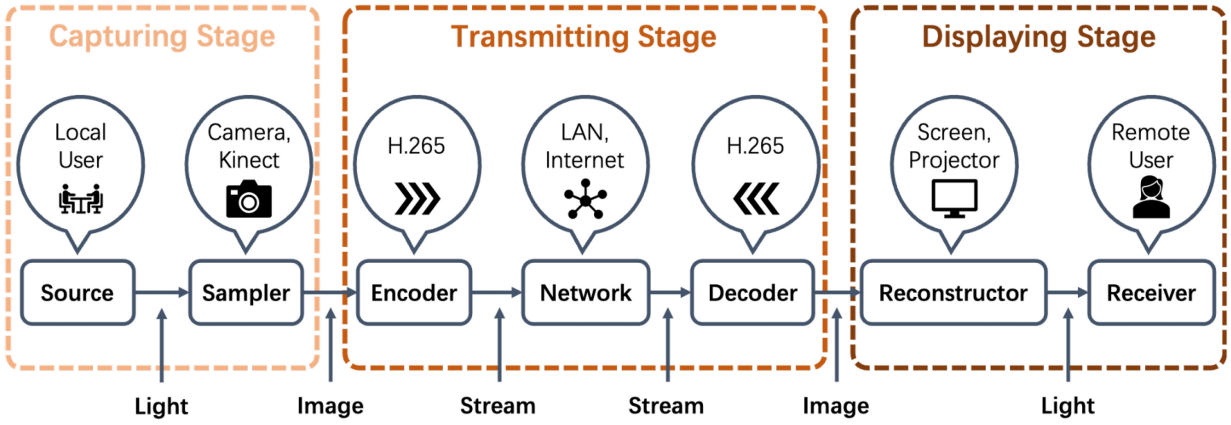


Fig. 3. Data flow of the visual *channel* can be divided into three stages: Capturing stage, Transmitting Stage and Displaying Stage.

information content that can be conveyed from source to destination. We observe that the limitations of many existing approaches, such as lack of non-verbal cues [8], deixis challenges [16], "Mona Lisa effect" [17] by video medium, *et al.*, are mostly caused by information loss. If there are several perfect *channels* that are able to bring complete sensory information to remote users in an unaware manner, people can hardly determine whether they are joining a local meeting or a video-mediated conference. Therefore, we choose *capacity* as the first metric to measure the information content successfully conveyed to the remote user end. After building an information content [19] model for each data modality, the second criterion is *cost*, including both the complexity of devices and the extra burden that a system brings to users during usage. Complex and heavy-weight devices result in high prices, and some extra modules, such as a head-mounted display, can significantly affect the user experience in the telepresence edge scenario.

Take the visual dimension as an example to describe the workflow of a *channel*. To facilitate our modeling, the vision subsystem is divided into three major stages: Sampling Stage, Transmitting Stage, and Reconstruction Stage as shown in Fig. 3. Information loss can occur in each stage. However, with the advances in coding algorithms and networking techniques, information loss in the coding and transmitting process can be neglected, so we mainly focus on the sampling and reconstruction stage.

We will also discuss the *channel* analysis for different modalities separately. It is worth mentioning that different from analysis in traditional computer networking systems which conducts machine-to-machine information transmission, the *channel* we put forward in this work convey information from user-machine-user, so to paying attention to the quality of user end.

Visual Dimension: Before evaluating the visual *channel*'s *capacity*, we require an information content model to represent complete visual information in an arbitrary scenario. In this work, we introduce Plenoptic Function [46] as the visual information model which includes all light rays from a particular viewing position.

$$P = P\big(\theta, \phi, V_x, V_y, V_z\big). \tag{2}$$

To understand this definition, we can imagine placing an ideal eye at every location (V_x, V_y, V_z) and recording the color value of the rays projecting into the pupil from every possible angle (θ, ϕ). Then the Plenoptic Function indicates a 5D light field that encodes all the image information within a scenario. In fact, by enumerating all possible locations and viewing angles, we can get a set of panoramic images that can be a representation of the 5D light field. An image captured by the camera can be regarded as a projection from the 5D vision space to a 2D plane. Meanwhile, the displaying stage can cause visual information loss as well.

In Information Theory, the information content of a message is determined by the amount of "surprise" or "uncertainties" conveyed by the message. According to Shannon's definition [19] and Plenoptic Function, the information content of a visual scene is represented by the following differential entropy.

$$H_v(X) = -\int_{V_x,V_y,V_z}\int_{\theta,\phi}\int_{r,g,b} p(x)\log p(x)\mathrm{d}x. \tag{3}$$

Then the information delivery rate is measured by the average mutual information between the source and the receiver.

$$I_v(X;Y) = H_v(X) - H_v(X|Y). \tag{4}$$

And the *channel capacity* is defined as the maximum $I(X;Y)$ respect to varying input distributions:

$$C \overset{def}{=} max_{p(x)}\{I_v(X;Y)\}. \tag{5}$$

From the above equations, we can find that the *channel capacity* is determined by the source information content $H(X)$ and residual information content $H(X|Y)$. We have discussed the source information content $H(X)$ before and $H(X|Y)$ can be described as the amount of "uncertainties" a remote user still holds after watching the images provided by the telepresence edge system. In other words, the *channel capacity* indicates the most possible "uncertainties" that can be eliminated.

In the video-capturing stage, a single camera can be regarded as a sampler from 5D Plenoptic Function to a 2D image plane. After this sampling, how many "uncertainties" still exist? Since "uncertainty" comes from the absence of visual information in the sampled image, we consider all light

rays that are out of the camera's field of view when calculating the residual information content.

$$H_v(X|Y) = -\int_{V_x,V_y,V_z}\int_{(\theta,\phi)\notin f}\int_{r,g,b} p(x)\log p(x)\mathrm{d}x. \quad (6)$$

In practice, the above integration is hard to calculate, therefore, we make some assumptions to simplify the calculation.

- First, the visual information in different regions is not of equal importance in a mobile officing scenario. The images containing people are the most we need, the background of a person provides little information. Under this assumption, we select the portrait region as our *target region*.
- Second, we assume that the information contained in *target region* follows an independent and identical distribution. Therefore, we apply the angle of view of cameras to approximate the coverage information content.

Thus we approximate the *channel capacity* of the capturing stage with the following expression.

$$\tilde{C}_c = \frac{N\alpha}{\pi}H(X), \quad (7)$$

where α denotes the angle of view that a camera can cover, N is the number of separate cameras around the target person (here we assume the view angles of these cameras are not overlapped), and $H(X)$ represents the source information content. Now, we have an approximation of the *channel capacity* for the image-capturing stage. Particularly, some multi-view stereoscopic techniques [8], [12] leverage several cameras to reconstruct the integrated 3D visual information of a person, in which case we set $C = H(X)$.

Next, we look at the displaying stage. Similar to that in the capturing stage, the *channel capacity* of the displaying stage C_d is measured by the difference between $H(X')$ and residual uncertainties.

$$\tilde{C}_d = H_v(X') - H_v(X'|Y). \quad (8)$$

$H(X')$ denotes the source information content in displaying stage. We assume $H(X) = H(X')$ to facilitate the following discussion and similar analysis can be conducted while $H(X) \neq H(X')$. Theoretically, perfect display equipment can show all visual information simultaneously, that is, reconstructing the complete light field of a Plenoptic Function. In practice, this can be extremely difficult and expensive and is not necessarily as well. We observe that human cannot sense all light rays in a scene simultaneously. Human eyes can be regarded as two parallel cameras which can only take pictures from one static viewpoint at a time, bringing us the opportunity to simplify the displaying problem. It is enough to offer users rendered images in real-time from their desired viewpoints. Some existing efforts have achieved this goal with different techniques, *e.g.*, autostereoscopic displays [11], [35], head-mounted displays (HMDs) [13], and the like. Other approaches only provide users with images from the limited (usually fixed) view of points which leads to "uncertainties". Thus, we propose to measure the residual by the differences in view angles between the displayed image and user desired one:

$$\tilde{C}_d = \frac{\pi-\beta}{\pi}H_v(X'), \quad (9)$$

$$\tilde{C} = \gamma\tilde{C}_d\left(\tilde{C}_c\right), \quad (10)$$

where β denotes the summarized view angle differences and γ is a penalty factor taking some other sources of information loss into account, such as lack of awareness of parallax, *etc.*

Auditory Dimension: Similar to the Plenoptic Function for vision, we first present a model to characterize the complete set of auditory information. In this work, we only consider the human voice since it is the dominating auditory information for a conference scenario. We observe that the human sound heard by another person is determined by three major factors, 1) the audio streams which encode verbal information, the timbre of the speaker, the volume, *etc.*, 2) the relative direction of the speaker which can be captured by an audience due to the *Binaural Effect*, 3) the distance between speaker and audience. Different from light rays that can come from any point in a scene, sound waves have discrete sources. Then we define an Auditory Function A, which encodes all auditory sources around an audience. The information content $H_a(X)$ are also defined as follows:

$$A = A(\theta,\phi,d), \quad (11)$$

$$H_a(X) = -\int_{\theta,\phi}\int_d\int_a p(x)\log p(x)\mathrm{d}x, \quad (12)$$

where a is the possible sample value of a sound source. According to the above definition, we can find that if we reconstruct every sound source with proper relative direction and distance to the audience in remote rooms, we can say that full auditory information has been conveyed to the remote users. In this way, we expect that when users close their eyes they can hardly determine whether the speaker is local or online. As the integration of $H_a(X)$ is difficult to calculate, we consider the audio information as three parts and evaluate them separately. In fact, recent advances in sound sampling and coding techniques are able to offer high-fidelity sound, however, few of them address the "uncertainties" in direction and distance which result in a decline in user experience.

Other Dimensions: For the mobile officing scenario, we assume that there is no significant directional information in odor. Besides, as meeting rooms can be regarded as enclosed spaces, we assume that the scent is identical among the whole room. In this case, the information content of olfactory information can be calculated as

$$H_o(X) = -\sum_o p(x)\log p(x). \quad (13)$$

According to the above definition, we can find that if we share the scent among all distributed meeting rooms, we can eliminate all "uncertainties" in the olfactory dimension.

The tactile sensation in a meeting scenario usually comes from the physical interaction between different persons, *e.g.*, shaking hands. While in a VCS, the users can be together in a virtual space where virtual interaction and tactile feedback [47] can be achieved by wearable devices. In this case, the information content model of the tactile sensation can be defined as follows:

$$H_t(X) = -\int_{V_x,V_y,V_z}\int_t p(x)\log p(x)\mathrm{d}x, \quad (14)$$

TABLE I
TECHNIQUES APPLIED IN DIFFERENT DIMENSIONS

Sensory Dimension	Capturing Stage	Displaying Stage
Visual	Single-Camera Multi-Camera Multi-View Stereo 3D Depth Camera [8]	Fixed-Viewpoint Dynamic-Viewpoint Retroreflective Multi-View [6] Cylindrical Display [10], [11] CAVE Display [45] Multi-View Lenticular Display [34] Head-Mounted Display Robotic [14]
Auditory	Single Microphone Microphone Array Vision-Assisted	Single Speaker Double Speaker Multi-Channel
Olfactory	Olfactory Sensation	Scent Generator
Tactile	Wearable Sensor Vision-Assisted Collision Detection	Haptic Feedback System
Gustatory		
Resource	Digital Material Distribution Mark Recording Touch Screen Deictic Gesture Perception AI-Based Analysis	Digital Material Display Image Blender 3D Interaction
Space		Relative Positioning System Global Coordinate System

where (V_x, V_y, V_z) denotes the 3D locations in virtual space and t indicates the strength of tactile. Then, for the tactile dimension, the *channel capacity* is determined by the number of contact points, including their location and haptic strength that we can convey among distributed sites.

The information content model of gustatory sensation is similar to that of scent. As it is negligible for a conference application, we don't put much attention to this dimension.

As described above, the shared resource dimension mainly includes three categories of information. Because the shared resources usually exist in the form of digital materials, we can directly encode and distribute them through the Internet without an information-capturing stage and then display them in the remote sites. To this end, we measure the *channel capacity* of this dimension by its capability of providing and sharing the above three categories of information.

Finally, let's look into the spatial awareness dimension. In a local meeting room, all participants can feel that they are in the same physical space. While in telepresence edge systems, we consider this problem from two aspects. First, a global coordinate system and thus a virtual space should be established. Second, we need to assign locations for different participants in the virtual space and make users aware of the shared global space through the reconstruction of visual information, auditory information, and the like. For the purpose of evaluating the awareness of space property, we propose to measure the *channel capacity* by evaluating whether a VCS successfully establishes and shares among all participants a global virtual coordinate system.

C. Cost Analysis

In our model, the *cost* of each *channel* includes two aspects, the complexity of devices and the extra burden that a system brings to users on usage. The first aspect corresponds to the building price of the system. Some solutions leverage only commercial devices without extra costs such as Web cameras or laptops, and others require dedicated equipment such as a projector ring or robotics. The second aspect measures to what extent the telepresence edge system changes users' habits in a mobile officing scenario, *e.g.*, wearing VR HMDs or haptic feedback gloves. According to the above two measurements, we roughly classify existing telepresence edge systems into 4 categories, that is, COTS devices, dedicated devices, high-weight devices, and wearable devices. Different scores are assigned to every category, the higher the *cost*, the lower the score.

IV. MODEL-BASED ANALYSIS AND PROTOTYPE DESIGN

We select 29 typical telepresence edge systems and analyze them based on the proposed model. Table I shows some typical techniques applied in existing solutions for each information modality. For *channel* measurement, we first select proper values for parameters in our model. We set $\alpha = \pi/3$ in visual dimension because many multi-view stereo systems can reconstruct high-quality 3D silhouettes with just three commercial cameras. Then, the parameter γ gives a 10% penalty (γ=0.9) to each of the following situations, such as lack of motion parallax or limited display scope. This is because according to the previous indicators, many work results are consistent in the same dimension, but in fact, there are still some differences or defects when compared with each other. For example, both displays show 3D images but the cylindrical shape display can only present one person's image at a time [10]. We hope to characterize these defects that are not dominant factors so that they can be distinguished in the final result without affecting the performance of the main factors, thus a small penalty parameter is artificially set for visual dimension. For audio, we treat audio stream, distance, and direction as equally important. For tactile information, we estimate the coverage of haptic feedback with respect to

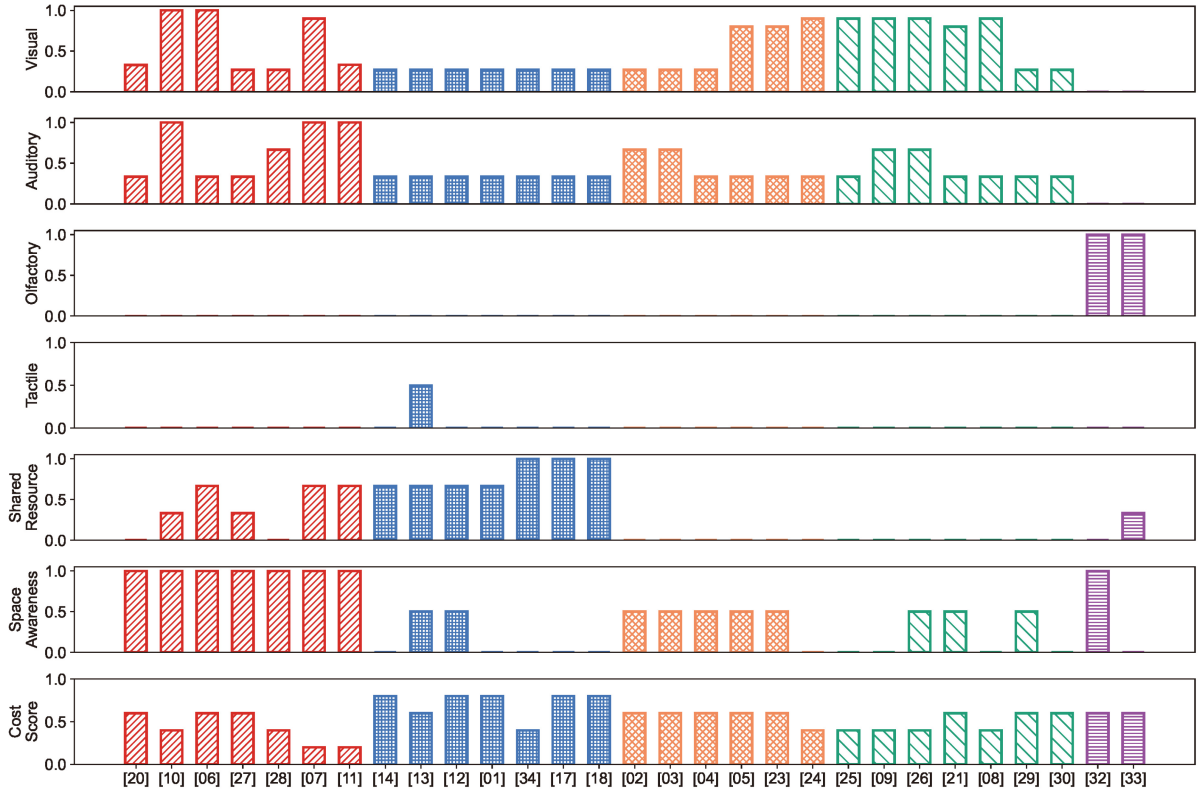


Fig. 4. The evaluation results of 29 typical telepresence edge systems according to our model show the trend and performance of emphasis in this field. Through K-means clustering, we classified similar works and they are Robots and HMDs that focus on spatial information (red); WYSIWIS approaches (blue); gaze-aware focusing (orange); 3D video and motion parallax providing (green); and olfactory aid (purple).

the whole human body. The gustatory dimension is ignored in this study because this modality is of little significance to general conference applications. As in our model, the *channel capacities* are all represented as an expression of their source information content, thus we normalize the results to [0, 1]. The normalized *capacity* can be regarded as a score for each dimension. The shared resource dimension is handled in a similar way as that in audio, three major categories of information are weighted equally. For measuring the awareness of space, we set the global coordinates system with a score of 1 and the relative positioning system with a score of 0.5. Finally, we empirically assign cost scores to different approaches according to their *cost*.

A. Insights and Guidelines

Note that the scores of different dimensions naturally form a feature vector, we conduct a K-means clustering of these methods using these feature vectors. As shown in Fig. 4, we find correlations among methods within each group. Approaches [13], [25], [40] in the first cluster attaches significant importance to awareness of space, which usually applies robots or HMDs in their systems, resulting in extra burden to users and relatively high cost. The second cluster includes most of the WYSIWIS [16], [18], [47] methods that focus on building a shared 2D space where images and other types of information are blended. In this way, rich information about shared resources can be conveyed. These methods generally leverage low-cost COTS devices. The third group of methods applies 2D images as well, but are aware of gaze direction [4], [5] through some relative positioning schemes. The fourth cluster pay more attention to capturing and reconstructing rich visual information such as 3D video [11], [30], [37]. Motion parallax is also supported by some of them. The final category of efforts mainly considers olfactory dimension [48], [49].

Based on the above observation, we conclude three main shortcomings of current VCSs with opportunities for future remote collaboration.

- The visual dimension is the main area where the existing work is focused. Many SOTA approaches require expensive and complex devices while supporting only a limited number of users, and like showing images of only one person at a time or resulting in extra burdens to users such as wearing shutter glasses or HMDs that bring uncomfortable feelings and inconvenience. Based on this insight, we should attach more importance to the validity of *cost* and information delivery.
- The information transmission of the auditory dimension mainly focuses on verbal content and seldom pays attention to the aspect of the direction and distance of sound, declining the authenticity of the conversation and the awareness of the shared space.
- Besides, other dimensions are being seriously neglected in current mobile officing. We believe that information from different modalities is complementary to constructing a good conference environment and improving the quality of UX. In addition, cross-modality information

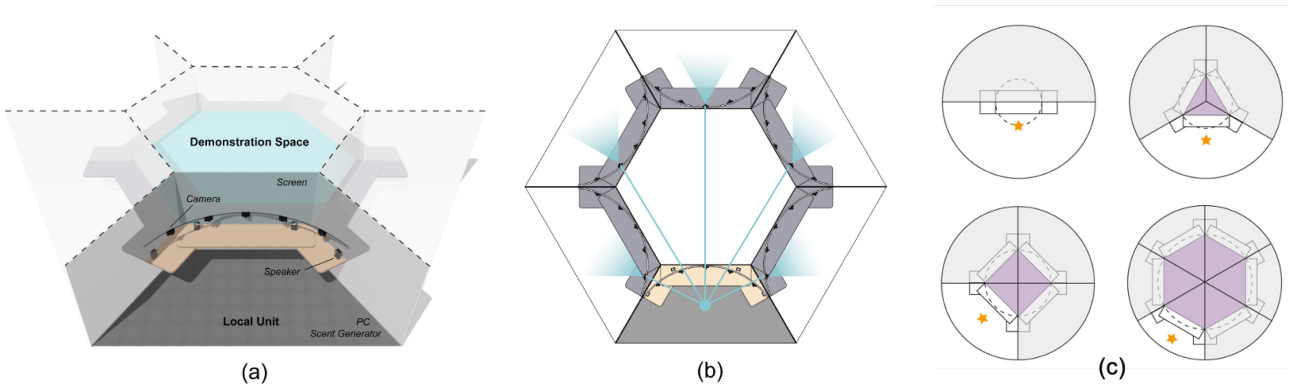


Fig. 5. The architecture of our prototype: (a) Hardware arrangement, including screen, cameras, microphone and speakers, scent generator, table, PC, etc. (b) Video streams selected for a local unit (c) Different layout when different number of parties join the same meeting room.

deduction and fusion could bring potential opportunities. For example, it is difficult to infer the direction and distance of a speaker by sound but can be easily obtained with the help of visual information.

To go further, providing a sense of ritual, coexistence, and free from constraint is of significant importance to remote collaboration. Also, there has been a significant trend to augment online applications with rich information from AI analyzers [50] such as auto-summary [22], talk tracking [21] and mood recognition. We believe this would be a good direction to raise the meeting efficiency and thus help people achieve a "beyond being there" solution.

B. Prototype Design

Based on the lessons and guidelines learned from the model-based analysis, our design takes video conference as the scenario and aims to have high UX quality and modest cost, which achieve high validity of multi-modality information transmission. We first present some *cost* restrictions:

- We only use COTS components in our prototype.
- We do not bring extra burdens to users or change their habits during the meeting process.
- The proposed system is not dedicated to "one-to-one" meetings but supports flexible multi-party conferences.

First, we build a global coordinate system and assign a region for each party. As illustrated in Fig. 5 (a), the shared conference space is a fusion of real meeting rooms and virtual ones corresponding to remote parties.

For visual dimension, we use commercial equipment to provide an immersive conference experience with dynamic viewpoints and multi-angle asynchronous pictures. We apply a CAVE-like display system [12], [45] consisting of three wall-size projection screens. The local user can "see" other participants through these screens. In this design, we only convey the segmented human images which are fused with the 3D virtual environment of the shared space. As shown in Fig. 5 (b), a ring of Web cameras is deployed in front of each local user to capture images of her/him from varying perspectives. Different video streams are sent to corresponding remote participants according to their relative positions in the global space. Mutual eye contact and other nonverbal information are preserved.

For the audio dimension, we collect every user's voice with two omnidirectional microphones mounted on the desk and track local users' locations through image analysis. The audio stream together with location information is distributed to all other participants. Then for each remote party, the sound is reconstructed and displayed according to the relative position between the speaker and listener in the global virtual space. In this case, the verbal content in audio, direction, and distance information are all preserved, and thus users can enjoy a similar audio experience just like they are in the same place.

In a conference scenario, directly capturing and sharing olfactory information among different parties is of little consequence, which sometimes can even lead to a worse situation, *e.g.*, sharing bad odor in one place can lead to an experience decline of all participants. Due to this consideration, our prototype determines the odor to be shared by an AI-based atmosphere analysis model. The atmosphere analysis model takes real-time video and audio as input and estimates the current atmosphere of the conference, then the proper scent is shared with a generator in each local room, refreshing the participants or making them calm.

For the shared resource dimension, we set the middle region of our global space as the demonstration area where images, videos, and 3D simulation models can be displayed. Users can put marks on as well as interact with these shared resources and all other participants can get the information.

We do not convey tactile and gustatory in this design as these two modalities contribute a few for communication tasks in a conference, and will bring extra burdens to users. Our design considered other novel features such as the flexible conference space. The virtual-physical fused conference space can support different numbers of parties joining a meeting simultaneously, which is shown in Fig. 5 (c).

We also evaluate our prototype design with the proposed model and the results are shown in Fig 6. We can find that our design achieves reasonable scores among nearly all dimensions which means it is able to achieve good performance at a modest cost.

TABLE II
SENTIMENT CLASSIFICATION PERFORMANCE WITH DIFFERENT KERNELS

Metrics	Kernels			
	Linear	**Polynomial**	**Sigmoid**	**Gaussian**
Accuracy	**83.43%**	79.01%	79.01%	82.87%
Balanced Accuracy	**77.15%**	74.71%	69.35%	76.77%
F1-Score (macro)	**78.32%**	76.85%	70.89%	78.27%
95% Confidence Interval of Accuracy	**[78.01%, 88.85%]**	[73.08%, 84.94%]	[73.08%, 84.94%]	[77.38%, 88.36%]

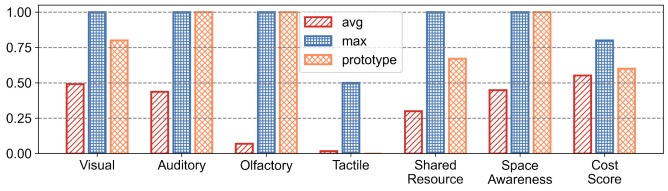


Fig. 6. The evaluation results of our prototype design according to the model compared with the maximum and average performances of 29 typical telepresence edge systems.

V. IMPLEMENTATION AND EXPERIMENT

A. System Implementation

For the visual part, three 3200mm x 2400mm curtains located on the two sides and in the front of the CAVE formed the interface. The angle between the side screen and the front screen can be adjusted within 90 to 180 degrees through articulation. As the number of attending parties changes, the angle of the screens on both sides and the structure of the conference table can be dynamically adjusted to achieve the seamless integration of multiple local units in the virtual space through coordinate mapping as shown in Fig. 5 (c). The virtual conference scene is pre-built on the Unreal Engine 5 platform, and the number of parties supported is 2-6 in consideration of the optimal meeting efficiency and user experience. In order to ensure the visual acquisition of the correct perspective and avoid the "Mona Lisa effect", we set up a ring of cameras at the height of 1250mm, which is flush with the observer's general sight height when sitting. Webcams are placed on tripods for real-time shooting with a resolution of 1080p/30fps. The number and angle of the webcams are determined by the number of attending parties, always one less than the number of parties. Although the cameras are in the user's field of view, the later experiment proved that it has no effect on the user's acquisition and transmission of information.

For the acquisition and playback of surround sound [51], we collect the sounds from each party and implement audio panning based on the Pydub package. Given the number of participants and their directions, audio streams can be panned respectively to achieve a spatial sense by using two speakers. In addition, the sound can be zoned during group discussions to automatically block the sound of other groups.

We also place a scent generator containing two different scents, mint and sandalwood. Sandalwood can relieve anxiety and settle people's central nervous system; mint can refresh the mind and eliminate sleepiness. After collecting the visual and auditory streams, we analyze multi-modality information to define the whole conference's atmosphere and automatically release the scent. The essential oil burets are connected to the Arduino UNO board through the atomization pieces, so that the category of scent release can be controlled. The detailed method will be described in Section V-B, and the final prototype is shown in Fig. 7(a)-(c).

B. Atmosphere Analysis and Experimental Results

In our system, the atmosphere during the conference is evaluated by recognizing the sentiments of participants, which is a cross-modality information processing that combines vision, auditory and olfactory dimensions. Our method takes video streams as input and features are extracted using a neural network structure named MIMAMO [52]. The obtained features are represented with the Arousal-Valence space [53], in which arousal represents the activeness of the emotion and valence indicates whether the emotion is positive or negative. We apply an SVM classifier to recognize the sentiment of users and thus determine the atmosphere of the whole conference.

To construct the train and test dataset, we collect and segment video clips of real-life video conferences from the Internet, and the final dataset includes 454 video segments in total. The dataset is randomly split into a train set that contains 60% data and a test set containing the rest 40%. The evaluation results are shown in Table II, according to which we find that the Linear kernel achieves the best performance than other kernels. In fact, Linear kernels behave similarly to Gaussian kernels, and there might be no significant difference in the overall implementation of the system, so we ultimately chose the former with simpler structures and faster computations.

C. User Study

The user study aims to qualitatively model the telepresence edge systems' performance and explore users' direct perceptions and feedback when receiving multidimensional information. We hope to gain a more in-depth understanding of the UX quality of our model-induced prototype and to prove the validity of our IT model.

Participants: We recruited 20 participants (8 women, 12 men, age range 21-52, M=28.15, SD=9.38) from our university. All our participants are familiar with the most common commercial cloud-based VCSs, such as Tencent Meeting and Zoom while also often participating in offline meetings.

Procedure: We first asked the participants to sign the consent and instruction form. Then, they were asked to hold a five-minute online meeting with five other people using

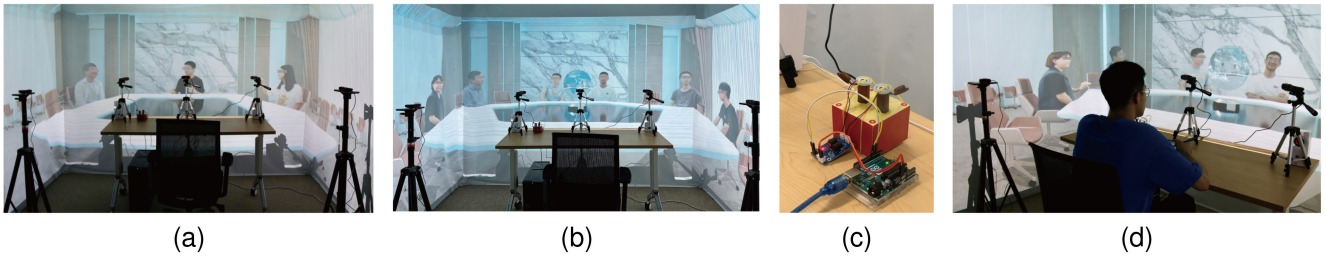


Fig. 7. Implementation and user study environment of our design: (a) Four-parties conference, (b) Six-parties conference, (c) Scent generator, (d) User study environment.

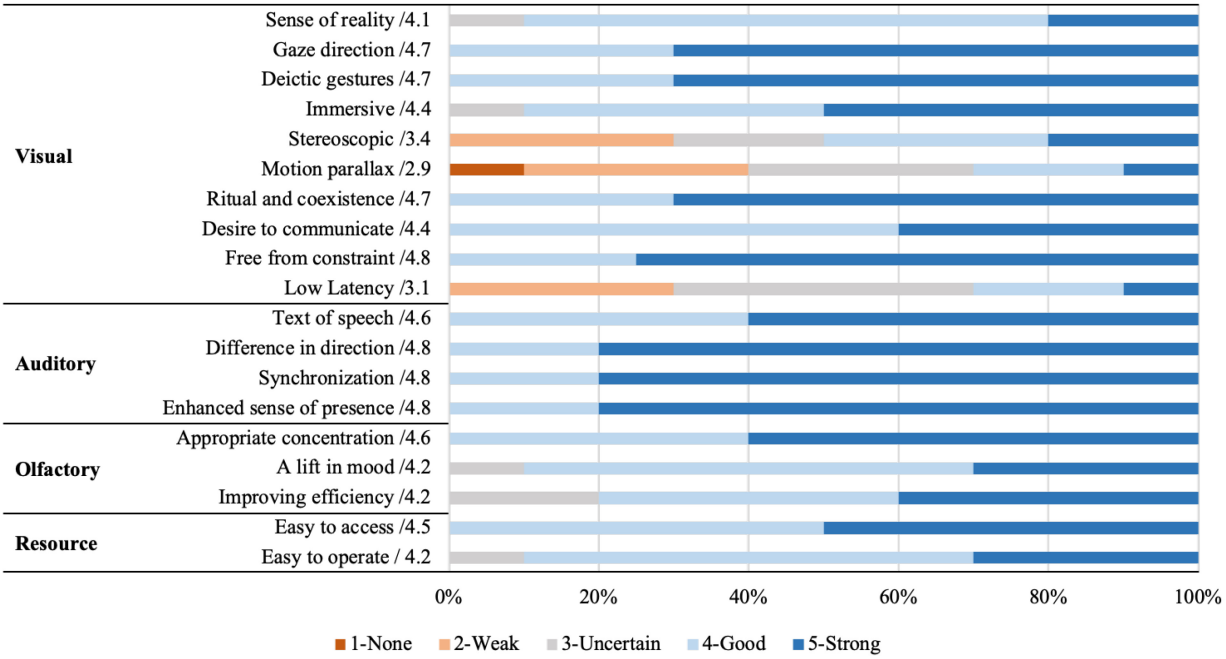


Fig. 8. The results of the user study, using more precise metrics of user experience from visual, auditory, olfactory, and resource dimensions.

Zoom, during which they switched on the camera and made specified eye contact, gestures, and conversations. The UX quality in this scenario will serve as a baseline for comparison. The next step is to get participants into the local CAVE to hold a conference with five remote parties shown in Fig. 7(d). This phase is a pre-test that participants can exit if they feel dizzy or uncomfortable. In fact, all participants completed the experiment without any adverse symptom feedback. After that, the scene switched to a four-party meeting. Surround sound and scent was added gradually so that the participants could feel the seamless integration of virtual, auditory, and olfactory senses in the conference room. In this scenario, they are also asked to make specified eye contact, gestures, and conversations just the same as in Zoom. For the last step, participants filled out our UX questionnaire, and the researchers conducted a ten-minute interview based on their feedback.

Questionnaire, analysis, and results: Our questionnaire is based on a 5-level Likert Scale and measures UX quality from visual, auditory, olfactory sensory, and resource dimensions including 19 criteria. These metrics are all obtained through our intensive analysis of the existing VCSs and theories, which are also related to our IT model and insights. The experimental results shown in Figure 8 basically meet our expectations. The model-induced prototype achieved more than 4 points in each sensory aspect and participants rated the overall experience at an average of 4.5 out of 5. Participants scored particularly high in free from constraint (4.8), gaze and gesture direction awareness, sense of ritual and coexistence (4.7), immersive, desire to communicate (4.4) as well as the perception of different directions of sound based on audio and video information (4.8). There is truly a lift in mood and concentration caused by odor release (4.6), and digital resources are easy to access (4.5).

We must admit that there are still shortcomings in some aspects, such as providing stereoscopic and low latency images, and achieving horizontal and vertical motion parallax. During the interview, some participants also mentioned the lack of tactile sense, the authenticity of light and character rendering. We will continue to improve them in the future.

VI. Conclusion and Future Works

In this work, we propose an Information Theory based model to characterize and evaluate the UX quality of the

existing telepresence edge approaches from a novel perspective. The information flow changes from traditional machine-machine to user-machine-user. We also discuss insights and heuristic guidelines derived from this descriptive model for future remote collaboration improvement. Then, based on the above modal-based analysis, we present a prototype design under certain cost restrictions. Finally, we implement the prototype and the validity of our method is validated by extensive experiments and user studies. Although the model-based analysis is inevitably biased, our study provides meaningful references for future telepresence edge system design. For our future work, we will first improve our model by making more accurate and detailed estimates. Second, we would like to improve our design, particularly paying attention to cross-modality information fusion and AI-based analysis.

REFERENCES

[1] "Zoom." Accessed: Sep. 2022. [Online]. Available: https://zoom.us/

[2] "Tencent meeting." Accessed: Sep. 2022. [Online]. Available: https://meeting.tencent.com

[3] K.-I. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita, "Multiparty videoconferencing at virtual social distance: MAJIC design," in *Proc. 5th ACM Conf. Comput. Support. Coop. Work*, 1994, pp. 385–393.

[4] A. Sellen, B. Buxton, and J. Arnott, "Using spatial cues to improve videoconferencing," in *Proc. 10th SIGCHI Conf. Human Factors Comput. Syst.*, 1992, pp. 651–652.

[5] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, "GAZE-2: Conveying eye contact in group video conferencing using eye-controlled camera direction," in *Proc. 20th SIGCHI Conf. Human Factors Comput. Syst.*, 2003, pp. 521–528.

[6] D. Nguyen and J. Canny, "MultiView: Spatially faithful group video conferencing," in *Proc. 23rd SIGCHI Conf. Human Factors Comput. Syst.*, 2005, pp. 799–808.

[7] D. T. Nguyen and J. Canny, "Multiview: Improving trust in group video conferencing through spatial faithfulness," in *Proc. 25th SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 1465–1474.

[8] Y. Zhang et al., "VirtualCube: An immersive 3D video communication system," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 5, pp. 2146–2156, May 2022.

[9] S. Orts-Escolano et al., "Holoportation: Virtual 3D teleportation in real-time," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, 2016, pp. 741–754.

[10] K. Kim, J. Bolton, A. Girouard, J. Cooperstock, and R. Vertegaal, "TeleHuman: Effects of 3D perspective on gaze and pose estimation with a life-size cylindrical telepresence pod," in *Proc. 30th SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 2531–2540.

[11] D. Gotsch, X. Zhang, T. Merritt, and R. Vertegaal, "TeleHuman2: A cylindrical light field teleconferencing system for life-size 3D human telepresence," in *Proc. 36th CHI Conf. Human Factors Comput. Syst.*, 2018, pp. 1–10.

[12] M. Gross et al., "Blue-c: A spatially immersive display and 3D video portal for telepresence," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 819–827, Jul. 2003.

[13] "Workrooms | VR for business meetings." Accessed: Sep. 2022. [Online]. Available: https://www.oculus.com/workrooms/

[14] N. P. Jouppi, S. Iyer, S. Thomas, and A. Slayden, "BiReality: Mutually-immersive telepresence," in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 860–867.

[15] C. Ishak, C. Neustaedter, D. Hawkins, J. Procyk, and M. Massimi, "Human proxies for remote university classroom attendance," in *Proc. 34th CHI Conf. Human Factors Comput. Syst.*, 2016, pp. 931–943.

[16] J. E. Grønbæk, B. Saatçi, C. F. Griggio, and C. N. Klokmose, "MirrorBlender: Supporting hybrid meetings with a malleable video-conferencing system," in *Proc. 39th CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–13.

[17] O. Morikawa and T. Maesako, "HyperMirror: Toward pleasant-to-use video mediated communication system," in *Proc. 7th ACM Conf. Comput. Support. Cooper. Work*, 1998, pp. 149–158.

[18] H. Ishii and M. Kobayashi, "ClearBoard: A seamless medium for shared drawing and conversation with eye contact," in *Proc. 10th SIGCHI Conf. Human Factors Comput. Syst.*, 1992, pp. 525–532.

[19] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[20] J. Hollan and S. Stornetta, "Beyond being there," in *Proc. 10th SIGCHI Conf. Human Factors Comput. Syst.*, 1992, pp. 119–125.

[21] S. Chandrasegaran, C. Bryan, H. Shidara, T.-Y. Chuang, and K.-L. Ma, "TalkTraces: Real-time capture and visualization of verbal content in meetings," in *Proc. 37th CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1–14.

[22] S. Samrose et al., "MeetingCoach: An intelligent dashboard for supporting effective & inclusive meetings," in *Proc. 39th CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–13.

[23] S.-K. Lee and J.-H. Kim, "Air-text: Air-writing and recognition system," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1267–1274. [Online]. Available: https://doi.org/10.1145/3474085.3475694

[24] S.-W. Fu, Y. Fan, Y. Hosseinkashi, J. Gupchup, and R. Cutler, "Improving meeting inclusiveness using speech interruption analysis," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 887–895. [Online]. Available: https://doi.org/10.1145/3503161.3548379

[25] S. J. Gibbs, C. Arapis, and C. J. Breiteneder, "TELEPORT—Towards immersive copresence," *Multimedia Syst.*, vol. 7, no. 3, pp. 214–221, 1999.

[26] L. Yang, M. Xu, T. Liu, L. Huo, and X. Gao, "TVFormer: Trajectory-guided visual quality assessment on 360° images with transformers," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 799–808. [Online]. Available: https://doi.org/10.1145/3503161.3547748

[27] Y. Jin, J. Liu, F. Wang, and S. Cui, "Where are you looking? A large-scale dataset of head and gaze behavior for 360-degree videos and a pilot study," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1025–1034. [Online]. Available: https://doi.org/10.1145/3503161.3548200

[28] O. Oyekoya, W. Steptoe, and A. Steed, "SphereAvatar: A situated display to represent a remote collaborator," in *Proc. 30th SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 2551–2560.

[29] Y. Pan, O. Oyekoya, and A. Steed, "A surround video capture and presentation system for preservation of eye-gaze in teleconferencing applications," *Presence*, vol. 24, no. 1, pp. 24–43, Feb. 2015.

[30] H. H. Baker et al., "Understanding performance in coliseum, an immersive videoconferencing system," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 1, no. 2, pp. 190–210, May 2005.

[31] A. Kulik et al., "C1x6: A stereoscopic six-user display for co-located collaboration in shared virtual environments," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 1–12, Dec. 2011.

[32] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The office of the future: A unified approach to image-based modeling and spatially immersive displays," in *Proc. 25th Annu. Conf. Comput. Graph. Interact. Techn.*, 1998, pp. 179–188.

[33] D. J. Sandin, T. Margolis, J. Ge, J. Girado, T. Peterka, and T. A. DeFanti, "The VarrierTM autostereoscopic virtual reality display," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 894–903, Jul. 2005.

[34] P. Lincoln, A. Nashel, A. Ilie, H. Towles, G. Welch, and H. Fuchs, "Multi-view lenticular display for group teleconferencing," in *Proc. 2nd Int. Conf. Immers. Telecommun.*, 2009, pp. 1–8.

[35] Y. Pan and A. Steed, "A gaze-preserving situated multiview telepresence system," in *Proc. 32nd SIGCHI Conf. Human Factors Comput. Syst.*, 2014, pp. 2173–2176.

[36] A. Jones et al., "Achieving eye contact in a one-to-many 3D video teleconferencing system," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 1–8, Jul. 2009.

[37] X. Zhang, S. Braley, C. Rubens, T. Merritt, and R. Vertegaal, "LightBee: A self-levitating light field display for hologrammatic telepresence," in *Proc. 37th CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1–10.

[38] J. E. Grønbæk, H. Korsgaard, M. G. Petersen, M. H. Birk, and P. G. Krogh, "Proxemic transitions: Designing shape-changing furniture for informal meetings," in *Proc. 35th CHI Conf. Human Factors Comput. Syst.*, 2017, pp. 7029–7041.

[39] M. K. Lee and L. Takayama, "'Now, i have a body': Uses and social norms for mobile remote presence in the workplace," in *Proc. 29th SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 33–42.

[40] "Test-driving willow garage's telepresence robot—CNET." Accessed: Sep. 2022. [Online]. Available: https://www.cnet.com/culture/test-driving-willow-garages-telepresence-robot/

[41] P. Lincoln, G. Welch, A. Nashel, A. Ilie, A. State, and H. Fuchs, "Animatronic shader lamps avatars," in *Proc. 8th IEEE Int. Symp. Mixed Augment. Real.*, 2009, pp. 27–33.

[42] D. Sakamoto, T. Kanda, T. Ono, H. Ishiguro, and N. Hagita, "Android as a telecommunication medium with a human-like presence," in *Proc. 2nd ACM/IEEE Int. Conf. Human-Robot Interact.*, 2007, pp. 193–200.
[43] M. Otsuki, K. Maruyama, H. Kuzuoka, and Y. Suzuki, "Effects of enhanced gaze presentation on gaze leading in remote collaborative physical tasks," in *Proc. 36th CHI Conf. Human Factors Comput. Syst.*, 2018, pp. 1–11.
[44] J. Yang, P. Sasikumar, H. Bai, A. Barde, G. Sörös, and M. Billinghurst, "The effects of spatial auditory and visual cues on mixed reality remote collaboration," *J. Multimodal User Interfaces*, vol. 14, no. 4, pp. 337–352, 2020.
[45] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti, "Surround-screen projection-based virtual reality: The design and implementation of the CAVE," in *Proc. 20th Annu. Conf. Comput. Graph. Interact. Techn.*, 1993, pp. 135–142.
[46] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proc. 22nd Annu. Conf. Comput. Graph. Interact. Techn.*, 1995, pp. 39–46.
[47] O. Morikawa, S. Hashimoto, T. Munakata, and J. Okunaka, "Embrace system for remote counseling," in *Proc. 8th Int. Conf. Multimodal Interfaces*, 2006, pp. 318–325.
[48] J. Brooks, S. Nagels, and P. Lopes, "Trigeminal-based temperature illusions," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376806
[49] D. Dmitrenko, E. Maggioni, G. Brianza, B. E. Holthausen, B. N. Walker, and M. Obrist, "CARoma therapy: Pleasant scents promote safer driving, better mood, and improved well-being in angry drivers," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2020, pp. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376176
[50] Y. Xiong and F. Quek, "Meeting room configuration and multiple camera calibration in meeting analysis," in *Proc. 7th Int. Conf. Multimodal Interfaces*, 2005, pp. 37–44.
[51] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Berlin, Germany: Springer, 2019.
[52] D. Deng, Z. Chen, Y. Zhou, and B. Shi, "MIMAMO net: Integrating micro-and macro-motion for video emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2621–2628.
[53] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, 2010.

Ruiqing Wang (Student Member, IEEE) is currently pursuing the dual master's degree in data science and information technology with the Global Innovation Exchange, Tsinghua University, Beijing, and the University of Washington, Seattle. Her research interests include human-computer interaction and AIoT.

Kebin Liu (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Shanghai Jiaotong University, China. He is a Research Associate Professor with the Global Innovation Exchange, Tsinghua University, Beijing, China. His research interests include Internet of Things, pervasive computing, and network diagnosis.

Ziyue Dang (Student Member, IEEE) received the B.E. degree in computer science and technology from Tsinghua University, Beijing, in 2023. He is currently pursuing the graduate degree with the Computer Science Department, University of California at Los Angeles, Los Angeles. His research interests include mobile computing, pervasive computing, and Internet of Things.

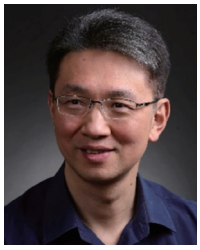
Xu Wang (Member, IEEE) received the B.E. and Ph.D. degrees in software engineering from Tsinghua University, Beijing, in 2015 and 2020, respectively, where he is a Research Assistant Professor with Global Innovation Exchange, Tsinghua University. His research interests include the industrial Internet, edge computing, and Internet of Things.

Fan Dang (Member, IEEE) received the B.E. and Ph.D. degrees in software engineering from Tsinghua University, Beijing, in 2013 and 2018, respectively, where he is a Research Assistant Professor with Global Innovation Exchange. His research interests include the industrial Internet, edge computing, and mobile security.

Yue Sun (Student Member, IEEE) received the bachelor's degree in electronic and computer engineering in 2021. He is currently pursuing the dual master's degree in data science and information technology with the Global Innovation Exchange, Tsinghua University. His research interest is 3-D computer vision.

Yuang Tong (Student Member, IEEE) received the B.E. degree from the Department of Automation, Tsinghua University, Beijing, in 2021. He is currently pursuing the dual master's degree with Global Innovation Exchange, Tsinghua University and the University of Washington, Seattle.

Haitian Zhao (Member, IEEE) received the Ph.D. degree from the School of Architecture, Tsinghua University, Beijing, in 2022, where he is currently a Postdoctoral Researcher with the Department of Automation. His research interests include smart building and AIoT.



Yunhao Liu (Fellow, IEEE) received the B.S. degree from the Department of Automation, Tsinghua University, and the M.S. and Ph.D. degrees in computer science and engineering from Michigan State University. He is the Chair Professor with Tsinghua University. He is a Fellow of ACM.