SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning

Zhipeng Li, Christoph Gebhardt, Yves Inglin, Nicolas Steck, Paul Streli, and Christian Holz
Department of Computer Science, ETH Zürich
Zurich, Switzerland

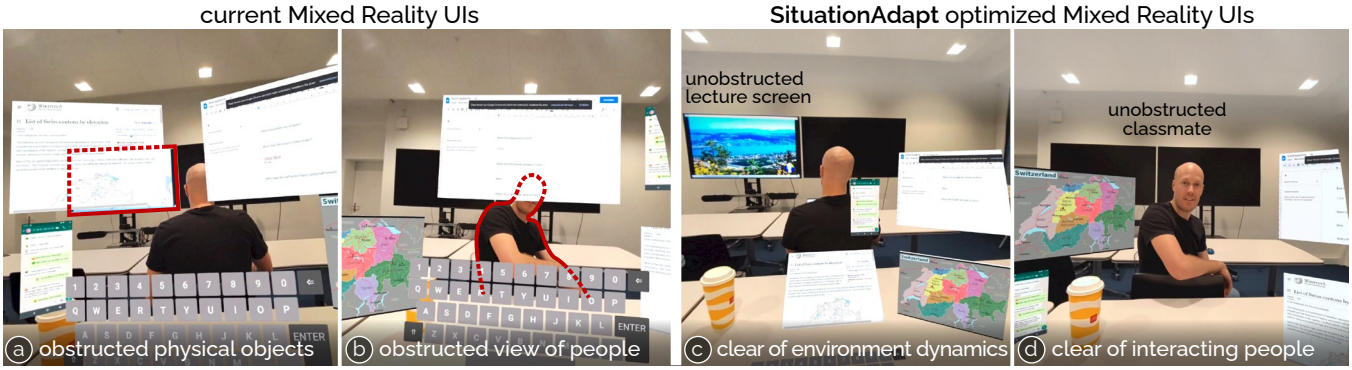


Figure 1: SituationAdapt is an optimization-based adaptive UI system that reconciles Mixed Reality layouts with shared real-world spaces. Previous layout adaptations do not consider the situational context, such as (a) if a shared display is on/off or (b) if a classmate is facing the user. Our computational pipeline identifies these and other characteristics and adapts Mixed Reality layouts with situational awareness, such that here (c) UIs stay clear of the video playback and (d) the talking classmate.

ABSTRACT

Mixed Reality is increasingly used in mobile settings beyond controlled home and office spaces. This mobility introduces the need for user interface layouts that adapt to varying contexts. However, existing adaptive systems are designed only for *static* environments. In this paper, we introduce *SituationAdapt*, a system that adjusts Mixed Reality UIs to real-world surroundings by considering environmental and social cues in shared settings. Our system consists of perception, reasoning, and optimization modules for UI adaptation. Our perception module identifies objects and individuals around the user, while our reasoning module leverages a Vision-and-Language Model to assess the placement of interactive UI elements. This ensures that adapted layouts do not obstruct relevant environmental cues or interfere with social norms. Our optimization module then generates Mixed Reality interfaces that account for these considerations as well as temporal constraints. For evaluation, we first validate our reasoning module's capability of assessing UI contexts in comparison to human expert users. In an online user study, we then establish SituationAdapt's capability of producing context-aware layouts for Mixed Reality, where it outperformed previous adaptive layout methods. We conclude with a series of applications and scenarios to demonstrate SituationAdapt's versatility.

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0628-8/24/10
https://doi.org/10.1145/3654777.3676470

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**; **Virtual reality**; **Interactive systems and tools**.

KEYWORDS

Mixed Reality, Adaptive User Interfaces, Large Language Models.

ACM Reference Format:
Zhipeng Li, Christoph Gebhardt, Yves Inglin, Nicolas Steck, Paul Streli, and Christian Holz. 2024. SituationAdapt: Contextual UI Optimization in Mixed Reality with Situation Awareness via LLM Reasoning. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3654777.3676470

1 INTRODUCTION

Mixed Reality (MR) devices are becoming increasingly more mobile, which indicates a future where they will be commonplace and can be used in shared public and private spaces. These can range from shared airplane [31] or train compartments, offices, and coffee shops to living rooms, kitchen areas, entire buildings [7] or public spaces [62]—similar to the environments where we commonly use smartphones, tablets, and laptops today.

Unlike user interfaces (UIs) on traditional screen devices, however, MR UIs transcend device boundaries; they can seamlessly blend into the user's physical surroundings and overlay parts of the real world. Adapting and reconciling virtual layouts with physical surroundings for MR use is a challenging task. Previous research has optimized MR UIs for proximity with semantically similar physical objects [8] or leveraged physical affordances of the user's surroundings to facilitate efficient interaction [9]. These adaptations have

so far focused on the real-world objects and surfaces within the user's reach inside their (personal) workspace, often assuming static environments during use.

In *shared* spaces, social norms become meaningful during interaction. Therefore, MR layouts must additionally conform to the social situations and dynamic environmental conditions that can take place in such environments. Previous studies highlight this need in their investigation of MR use in shared spaces (e.g., [21, 38, 44, 60]). The results of these studies indicate that users find it crucial for MR UI layouts to consider factors such as the functionality of objects in their surroundings, the social appropriateness of element placement, the effects of UI positioning on health & safety, and maintaining the visual appeal of the physical environment.

In this paper, we propose *SituationAdapt*, a system that optimizes MR layouts for situational social and environmental factors. Our system consists of perception, reasoning, and optimization modules to reconcile adapted MR UIs with real-world environments and conform to social norms and dynamic conditions.

Adapting UIs to Shared Real-World Settings

Figure 1 illustrates the challenge of situation-aware UI adaptation at using a lecture scenario. While a UI element can be suitably positioned in front of a classmate as he faces away from the MR user (Figure 1c), placing the same widget in front of his face as he faces or even interacts with the MR user is intrusive, as it not just impedes personal communication but also renders direct interaction with the MR UI inappropriate (Figure 1b). Likewise, UI elements may be placed in front of a physical screen, since they do not obstruct any information (Figure 1d). When the screen comes on, however, the virtual element occludes potentially meaningful content (Figure 1a).

SituationAdapt reconciles the layout of virtual UI elements with real-world conditions to ensure appropriate placement using the three modules of our system. This avoids intrusiveness and maintains considerate functionality in dynamic environments.

Our *perception module* identifies objects and people in the physical environment through a real-time object detection network while simultaneously reconstructing a 3D map of the user's surroundings. The module then segments identified objects and people from the 3D map to extract them as input into our optimization scheme.

Our *reasoning module* leverages a Vision-and-Language Model (VLM) to evaluate the potential placement of UI elements within a shared social space. Based on prior research, we designed a prompt to consider factors such as functionality, aesthetics, social acceptability, and health & safety. Because observing a UI element that occludes part of a shared space has different implications for these factors than a user's direct interactions with that UI element, we separately query the VLM for *overlay suitability* and *interaction suitability*. From the VLM response, we extract ratings to inform a goodness function for UI element placement that considers relevant environmental cues as well as social norms.

Finally, our *optimization module* processes the 3D bounding boxes of objects and people in the physical environment and the associated suitability ratings for overlaying content for display or interaction. From these inputs, the module generates layouts of MR UIs that account for environmental and social aspects of shared spaces. We propose two novel optimization terms for interactive MR adaptation that model the suitability for overlaying and interaction. Integrated into our real-time system, these terms optimize MR UIs for suitable viewing and interaction given the current shared physical environment.

We evaluate the efficacy of SituationAdapt in two studies: an online survey to evaluate our reasoning module and an in-situ user evaluation to evaluate our end-to-end system. In the online survey, we validate if the underlying VLM of our reasoning module judges the context of shared spaces similar to pre-screened, experienced MR users. We collected ratings from 42 participants and 42 VLM instances, evaluating 64 areas of interest within 18 diverse scenarios. The results of the survey indicate that, across scenarios, VLMs achieved comparable ratings to participants for both, overlay and interaction suitability.

We then conducted a user study to compare SituationAdapt's optimized layouts with those of two representative baseline methods that do not account for shared spaces. Participants perceived SituationAdapt's MR layouts to more suitably overlay UI elements onto the physical environment and position them more appropriately for interaction within the context of a shared social space. Participants also expressed a strong preference for the layouts generated by SituationAdapt compared to those from baseline methods. Finally, we demonstrate SituationAdapt's applicability across two scenarios within diverse shared spaces.

Contributions

We make the following contributions in this paper.

- an optimization-based end-to-end system that considers aspects of MR use in shared spaces in the optimization of MR layouts through an VLM-based reasoning component. Our approach can adapt UI element placements while taking into account their impact on, for instance, occluding real-world objects' functions, social appropriateness, health & safety, and the aesthetic appeal of the surroundings.
- a crowd-sourced survey study ($N = 42$) that demonstrated that our VLM-based reasoning module judges the context of shared spaces not different than experienced MR users.
- an empirical study that compared SituationAdapt to two baseline approaches ($N = 12$), showing that our approach generated layouts that participants preferred and rated more appropriate for shared spaces than the baseline layouts.
- two proof-of-concept scenarios that integrate our system to adapt MR layouts to the situational context of a shared space.

2 RELATED WORK

SituationAdapt is related to Mixed Reality usage in shared settings, adaptive layout systems for Mixed Reality, and the use of large language models in HCI.

2.1 Mixed Reality in shared spaces

Researchers have been exploring the effect of environmental and social dynamics of shared spaces on the use of MR devices [21] and interaction in MR. Transportation settings have been studied in depth [35, 41, 42], where shared surroundings demand socially acceptable and safe interaction [60], especially given the lack of space for expansive input [31]. Medeiros et al. studied the layout

of MR interfaces in shared transit contexts, including vehicles and trains, and identified important aspects for using VR in shared spaces: social etiquette, spatial affordance, and safety [44].

Other works have considered multiple users and bystanders within shared environments, such as for collaboration scenarios with multiple MR users [38] or individual MR users and projected augmented reality [22]. O'Hagan et al. explored the MR interactions with bystanders, reporting the need for socially intelligent bystander awareness systems [48].

While there are multiple factors influencing the experience of MR users and bystanders in shared spaces, it is hard to comprehensively model them in a computational manner. Our work leverages the reasoning capabilities of modern VLMs to understand the context of shared spaces and integrate inferred contextual information into an optimization scheme.

2.2 Adaptive Mixed Reality Interfaces

Prior research has explored adapting MR interfaces to various contextual factors including the user or their state, the task, as well as the physical environment.

One essential focus of MR adaptive user interfaces is environment-driven adaptation [15, 25, 26, 47]. Employing geometry-based approaches, researchers have suggested aligning virtual contents with the physical surroundings (e.g., Flare [18], Optispace [16], TapID [45], TapLight [54]). Lages and Bowman dynamically adapted virtual elements to physical windows and walls when the user was walking [34]. Qian et al.'s decision tree-based strategy adapted AR interfaces to new environments while keeping the semantic relationships between virtual and physical elements from the previous layout [50]. Kari et al.'s TransforMR method detected people and dynamic elements in MR scenes and substituting them with alternative avatars or objects through diminishing, thereby imbuing the physically plausible behavior of the original objects onto the synthetic replacements [30]. Asynchronous Reality dynamically diminished real-world objects to preserve the impression of the user's surroundings at one point in time when their state changed [17]. SemanticAdapt included the semantic relationship between virtual and physical objects with other factors such as temporal consistency, occlusion, and proposed an integer-programming-based optimization approach to obtain the adaptive interface [8]. Our previous UI adaptation method InteractionAdapt [9] additionally focused layout optimization on situated affordances such as physical surfaces and obstacles with empirically quantified benefits for interaction [10, 39] to provide passive haptic feedback and rest for optimized MR interaction during prolonged tasks between within-reach and far-away objects while accounting for physical obstacles that prevent input.

Other approaches investigated the adaptation of MR interfaces to the user's state [3, 37, 57]. Gebhardt et al. learned to display labels of virtual elements based on users' gaze interactions with the VR environment [19]. Lindlbauer et al. optimized virtual elements' visibility, level of detail and placement based on the estimation of users' cognitive load from pupil dilation [36]. Evangelista Belo et al. [14] and Montano Murillo et al. [46] further optimized virtual interfaces for ergonomics with rule-based estimation [40].

Newer work proposed a Pareto-optimal method to achieve a balance between competing objectives for MR UI adaptation [28] or introduced a tool to help researchers design new MR interfaces in various contexts based on previously collected MR UIs [11].

While research on adaptive MR interfaces explored numerous factors and settings, we are the first to adapt MR layouts to shared spaces considering factors such as 'social acceptability' and 'health & safety'. Our end-to-end system recognizes relevant cues in shared social settings and can optimize a MR UI accordingly.

2.3 LLMs in HCI

Recent advancements in Large Language Models (LLMs) have created widespread excitement across research disciplines, exploring their potential application to various tasks. In HCI, research has explored LLMs for tasks such as writing [12, 20], learning [4, 33], and programming [6, 49, 56]. Other works explored using LLMs to facilitate information retrieval [27], manage information with multilevel abstraction [55] and synthesize scholarly literature [29].

Most similar to our work, is research that uses LLMs to simulate participants of a user study. Hämäläinen et al. utilized GPT3 to generate open-ended responses about video game experiences and found that the LLM produced answers comparable to those of human participants [24]. Schmidt et al. found out that one might obtain artificial answers when using LLMs to simulate survey participants, but also highlight that LLMs give unanticipated responses that offer new insights and help to discover pitfalls in the survey design [53]. To validate LLM responses, they suggest to combine small-scale user studies with large-scale user simulation.

Following their suggestion, we validate the feasibility of our approach of utilizing a VLM to rate the suitability of placing virtual elements in shared social spaces with an online survey. In this evaluation, we compared VLM responses to those of experienced MR users in terms of understanding the context of shared spaces.

3 ADAPTIVE MR FOR SHARED SPACES

We define the factors to consider when developing adaptive MR layout approaches for shared spaces. By reviewing the aspects that previous studies consistently highlighted as crucial, we derive the following four key factors.

- **F** Functionality: UI elements hinder the functionality of a physical object (e.g., cup, laptop, display) [38, 44].
- **A** Aesthetics: UI elements impair the visual appeal of the physical surroundings [44].
- **S** Social acceptability: looking at or directly interacting with UI elements is considered socially inappropriate by bystanders [21, 43, 44, 60].
- **H** Health & Safety: UI elements occlude safety critical information or lead to sanitation issues during interaction [60, 61].

Furthermore, we respect that whether a user is soly observing a UI element or directly interacting with it can impact the FASH factors differently. For instance, while it may be socially acceptable for a user to glance at the map widget in Figure 1, direct interaction with it could be inappropriate, as it might distract other students attending the lecture. Similarly, placing a widget above the back of a passenger's head on a bus is suitable for observation but may be socially inappropriate for interaction, as it could lead to physical contact with the person's head. Therefore, we model suitability using two distinct scores: one for when a UI widget is being observed

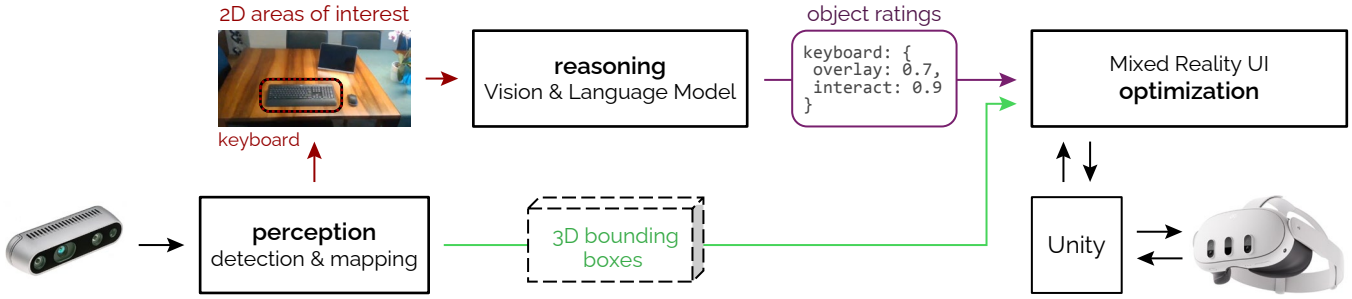


Figure 2: Schematic overview of SituationAdapt's system. Our perception module recognizes 2D areas of interest in the environment and computes 3D bounding boxes of the respective objects. Our reasoning module takes the areas as input and leverages a VLM to rate their overlay- and interaction suitability. Unity then assigns these ratings to the respective 3D bounding boxes and our optimization module adapts MR UIs accordingly.

(overlay suitability) and another for when it is being interacted with (interaction suitability).

To address these differences, we define *Overlay suitability* as the cumulative appropriateness of the **FASH** factors when a UI widget is being looked at and *Interaction suitability* as the cumulative appropriateness of the **FASH** factors when a UI widget is interacted with. We use these scores define the output of the VLM. With this formulation, the VLM can balance the impact of potentially conflicting **FASH** factors on placement suitability. This approach is more robust than treating each FASH factor as a separate objective term in an optimization scheme and relying on weight tuning to balance conflicting factors.

4 METHOD

SituationAdapt adjusts MR UIs to real-world conditions by considering social cues in shared settings. Figure 2 provides an overview of our system: a perception module recognizes objects and people around the user and fits 3D bounding boxes around them. Our reasoning module leverages a VLM to evaluate the suitability of scene locations to accommodate UI elements for display and/or interaction, ensuring that widgets do not obstruct relevant real-world cues or interfere with social norms. Our optimization scheme then uses these ratings as well as the 3D bounding boxes of the respective objects and people as input and generates MR interfaces that account for these aspects.

Below, we explain the operation of our modules. First, we discuss the functionalities and mechanisms of the perception module and the reasoning module. Finally, we detail the formulation of our optimization scheme.

4.1 Perception of surroundings

The perception module receives RGBD frames as input and provides semantically annotated 2D- and 3D bounding boxes of areas of interest as output. Areas of interest characterize the objects and people that were found in the real-world surroundings of the MR user, defined by the typical categories recognized by real-time object detection networks.

Modern MR headsets, such as Meta Quest 3, posses sophisticated inside-out tracking capabilities that can track the physical environment and even the dynamic user body. Recent developments indicate that these headsets will soon also have the capability to understand the 3D space around the user [52]. While these advancements already exist or are within reach, SDKs of current MR headsets do not make them available for developers. For this purpose, we developed a custom perception module (Section 5.1).

4.2 Reasoning about placement suitability

The reasoning module takes RGB images annotated with the areas of interest as input (Figure 4 illustrates examples of such images). For these images, we then query a VLM to rate the overlay- and interaction suitability for hypothetical UI elements positioned within box on a scale from from 1 ('unsuitable') to 5 ('suitable'). We start this evaluation by setting the context of the VLM, explaining what we mean with overlay- and interaction suitability of Mixed Reality UIs. We further prime it with the factors we derived to be important in the context of using Mixed Reality in shared spaces (Section 3). The comprehensive context prompt is detailed in Appendix A.

As initial tests revealed discrepancies between the VLM's ratings and user ratings, we have incorporated previously user-rated images and their respective ratings into the context of the VLM. More precisely, for each designated area within one of the user-rated images, we prompt the VLM with the median and the standard deviation of the ratings of a group of users. Using this context, we then query the VLM to rate overlay- and interaction suitability of a previously unseen image. Our tests have shown that this process increases the model's understanding of how users would rate situations and helps the VLM to align its ratings with those of users.

Finally, we query overlay- and interaction suitability for the areas of an unseen image with the following prompt: "Please rate the suitability of overlaying/directly interacting with a virtual UI element on each area in this image. The acquired ratings are forwarded to Unity and the optimization module.

4.3 Optimizing the MR UI layout

We base our optimization module on the AUIT toolkit [15]. The general form of the objective function of AUIT is defined as

$$Q = \sum_{i=1}^{V} \sum_{j=1}^{O} w_{ij} c_{ij}(\mathbf{x}) \tag{1}$$

where V is the set of virtual elements and O is the set of objectives, both accompanied by corresponding weights w and cost functions c. $\mathbf{x}$ is the decision vector comprising configuration parameters for all UI elements, optimized to minimize Q. In our optimization scheme, we utilize five pre-defined objective terms of AUIT: Occlusion, Look towards, Distance, Field of view, Constant view size (for details see the original paper).

To generate MR layouts that are sensitive to situations in shared spaces, we propose two new terms to model *overlaying suitability* and *interaction suitability*. Both terms take the detected 3D bounding boxes and the normalized 5-point suitability ratings (scaled between 0 and 1) as input. In contrast to the occlusion term in AUIT, which models the appropriateness of UI widgets being occluded by other UI widgets or physical objects, our terms consider the occlusion of real-world objects and people by virtual content during display and interaction, enabling situation-aware MR UIs.

To compute the *overlaying suitability* cost function, we rasterize each virtual element at an equal interval and cast a set of rays R from the users' point of view to each point within the grid. For each ray r, we then obtain a set of hit points $H(r)$ that constitutes the positions where the ray hit a 3D bounding box. Based on these sets, we can now compute the cost function $c_{v,over}$ for overlaying a virtual element v as,

$$c_{v,over} = \sum_{r}^{R} \sum_{h}^{H(r)} p_b e^{-5d_h}, \qquad (2)$$

$$d_h = \frac{||h - c_b||}{0.5 d_b}$$

where h is a hit point, c_b the center of the bounding box it hit, d_b the length of the box's diagonal, and d_h the respective normalized distance. We employ an exponential function to implement a higher penalty when the hit point is close to the center of the bounding box. The term p_b is the penalty of overlaying a bounding box b and is calculated as,

$$p_b = \begin{cases} 0.5 - o_b, & o_b \leq 0.5 \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

where o_b is the suitability score for overlaying the bounding box b. The term penalizes unsuitable boxes ($o_b \leq 0.5$), considering all others as suitable by default.

Similarly, we adopt the same grid-based ray casting procedure to compute the cost function for *interaction suitability* as

$$c_{v,inter} = \sum_{r}^{R} \sum_{h}^{H(r)} f_v (0.5 - i_b) e^{-5d_h} \qquad (4)$$

where i_b is the interaction suitability score of bounding box b. f_v represents how frequently a virtual element v is interacted with and hence needs to be penalized more in the context of this cost term (similar to respective terms in [8, 9, 36]). Intuitively, this term encourages placing virtual element over physical bounding boxes which are suitable for interaction by introducing a negative penalty when i_b is larger than 0.5.

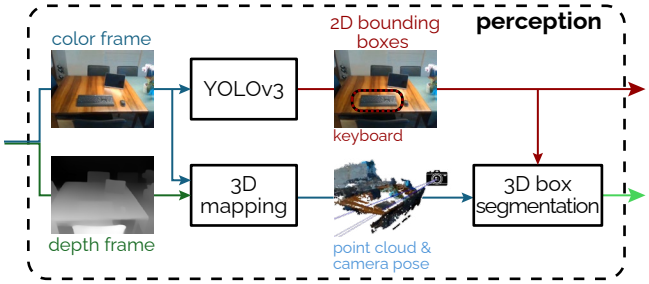


Figure 3: Our implementation of the perception module. Based on color- and depth frames of an RGBD camera, a 3D mapping stage reconstructs the camera position and the surroundings of the user as point cloud. An object detection node computes semantically annotated 2D bounding boxes. The last stage segments 3D bounding boxes based on the 2D ares, the point cloud and the camera position.

5 IMPLEMENTATION

We now outline the implementation of each of SituationAdapt's modules. Websockets facilitate the communication between them. Our entire pipeline runs on an Intel Core i7-12700K with a NVIDIA GeForce GTX 1050 Ti and 32 GB of RAM.

5.1 Perception module

The perception module aims to identify areas of interest as 2D- and 3D bounding boxes, serving as input for the reasoning- and optimization module. Our system utilizes the headset's inside-out tracking to maintain accurate positioning within the MR environment. We transform bounding boxes from our perception module to Unity using a manually specified transformation matrix. As our system adapts MR UI layouts at a situational change of a shared space, the perception module is manually triggered when such a change happens. We implemented the module within the Robot Operating System (ROS) where we ran separate ROS nodes for its three stages: 3D mapping, object detection, and 3D bounding box segmentation (see Figure 3 for an overview). Depth and color frames of an Intel RealSense D435 RGB-depth camera serve as input to the module. In the following, each stage is briefly explained.

5.1.1 3D mapping. We utilize the RTAB-Map implementation of the Simultaneous Localization And Mapping (SLAM) algorithm [2] to fuse RGB- and depth frames into a global 3D map of the surroundings. The resulting point cloud and camera position are forwarded to the 3D bounding box segmentation stage.

5.1.2 Object detection. We use YOLOv3 [51] to detect objects and people in the scene. It takes the color image as input and outputs a category, confidence, and bounding box for each detected object or person. The annotated 2D bounding boxes are forwarded to the reasoning module as well as the 3D bounding box segmentation.

5.1.3 3D bounding box segmentation. In this stage, we reproject the corners of the annotated 2D bounding boxes into the mapped 3D scene to create a frustum. This frustum is then transformed from camera to world coordinates, and its signed distance function is computed for point selection within it. Points not visible due to

occlusion are removed using the hidden point removal algorithm [32]. Finally, the DBSCAN algorithm [13] clusters the frustum's point cloud, a bounding box is built around the largest cluster, and points from other clusters are eliminated. Should a previously identified bounding box closely match the new one, it is replaced by the updated version. Conversely, if no similar bounding boxes are found, the new detection is incorporated into the scene as a separate entity. For each frame, all recognized bounding boxes are transmitted to Unity where a transformation is performed to convert them into Unity's coordinate system.

5.2 Reasoning module

We utilize the GPT4 Vision 2024-02-15-Preview model of Azure OpenAI as our VLM and access it via its Python API. As Azure AI Services lack the capability to fine-tune models through direct training on image data, we employ few-shot learning to provide our VLM with information about previously rated scenarios. This involves integrating example images and corresponding ratings, as described in Section 4.2, into its context prompt (see Appendix A for the specific prompt). We prompt the VLM to provide its answer in the format: Area <area index>: <score>, <reason>. The acquired ratings are transferred into Unity and assigned as properties to the respective 3D bounding boxes.

5.3 Unity & optimization module

We implement our system for the Meta Quest 3 using Unity 2021. To implement our MR UI optimization module, we leveraged AUIT [15], a toolkit to create adaptive Mixed-Reality applications. The toolkit interacts directly with Unity, utilizing Unity GameObjects and properties as input to its optimization.

6 REASONING VALIDATION

Our pipeline is built on the hypothesis that SituationAdapt can adequately understand the situational context of a shared space. To evaluate this assumption, we conducted an online survey to compare the judgment of different situations in shared social spaces of SituationAdapt with those of experienced MR users.

6.1 Survey design

Our survey sought to learn how SituationAdapt and experienced MR users judge the suitability of overlaying and directly interacting with virtual UIs in various scenarios and shared social spaces. In instances where parts of these scenes were deemed unsuitable for either, we further tried to discern which of factors we identified as critical (**FASH**) underlies the judgment. Thus, prior to starting the survey, we explained the suitability terms and the factors to participants. In addition, we showed two videos displaying the first-person view of a MR user in a share space. After the introduction, participants continued answering demographic questions before starting with the main part of the survey.

Scenarios. The main part of our survey consisted of 18 scenarios participants had to judge. Each scenario is a photo taken from first-person view of a hypothetical MR user. In each photo, we manually designed bounding boxes to create challenging scenarios for the VLM to analyze, following this rule: Placing a widget within the bounding box must affect one or more FASH factors (e.g., occluding or being near a person the user is talking to, blocking an important safety-related sign). Participants had to rate these areas for their overlay- and interaction suitability. We selected these scenarios to capture a wide variety of situations , including typical shared spaces (restaurant, airplane, home, office), social contexts (alone, with friends, strangers) and tasks (recreation, work). Figure 4 shows three scenarios presented in our survey with the respective areas (illustrated through bounding boxes) participants had to judge.

Questions. For each of the highlighted areas of a scenario, participants had to answer four questions. First, they were asked to rate the suitability of overlaying a virtual UI element on each area (*QO:* "Please rate the suitability of overlaying a virtual UI element on each area in a Mixed Reality experience"). Second, they should rate the suitability of directly interacting with envisioned virtual UI elements that were to be positioned in each area (*QI:* "Please rate the suitability of directly interacting with virtual UI elements displayed in each area. Note: All virtual elements are positioned within your arm's reach. If a virtual element covers a physical object, interacting with it means physically touching that object."). Responses to both questions were recorded using a 5-point Likert scale, with options ranging from "Unsuitable" to "Suitable" (1: "Unsuitable", 2: "Somewhat unsuitable", 3: "Neutral", 4: "Somewhat suitable", 5: "Suitable"). If participants selected 'unsuitable' or 'somewhat unsuitable' for an area in either question, they were asked to provide the reason (*QR-O/I:* "If you selected 'unsuitable' or 'somewhat unsuitable', please select the primary reason for your choice."). The response options corresponded to the underlying factors outlined in the survey's introduction (Functionality, Social Acceptability, Health & Safety, Aesthetics, Other with a text field to specify it).

6.2 Participants

We recruited 50 participants (16 female, 34 male), ages 22–50 (M=32, SD=9.1) from an online crowd-sourcing platform. To guarantee a certain level of VR experience among participants, we screened them to ensure they used a VR device at least 6 times a month. Of those, 13 participants reported using VR more than 15 times a month, 7 participants used it 11–15 times, and the remaining participants used it more than 6 times per month. Participants also reported their frequency of using direct touch to interact in Mixed Reality: 3 participants reported daily use, 13 mentioned using it several times a week, 18 indicated they used it several times a month, and the remaining participants used it less frequently. Participants completed the survey in 45 min and received £6 as a gratuity.

We excluded participants that answered one third of our control questions wrong (more than 6 out of 20 control questions) as well as participants that gave extreme extreme median responses (1: "Unsuitable" or 5: "Suitable") with a standard deviation lower than one across all areas and images. Consequently, the data from 42 participants (PTPS) were used in the analysis.

6.3 Generated suitability ratings

To generate results with SituationAdapt, we employed the identical scenarios and areas as for participants, and used the requests outlined in Section 4.2 to generate results with our perception module. To ensure a matching sample size, we produced ratings from 42

Figure 4: Our survey covered these and other scenarios. Participants rated the overlay and interaction suitability for each area.

distinct VLM instances, aligning the quantity of ratings between participants and SituationAdapt. We split the scenarios in training- and test set and add the training set (9 out of 18) to the context of VLM instances following the process described in Section 4.2. We then generated ratings for each question across the unseen scenarios, resulting in ratings from 42 VLM instances (VLMs) across 9 scenarios with 3 or 4 areas each. For our analysis, this yielded a total of 1,344 ratings per condition (VLMs and PTPs).

6.4 Results

The goal of our analysis was to determine if SituationAdapt assesses overlay and interaction suitability of social scenarios similar to the population of experienced MR users. Hence, we postulate the following null hypothesis:

H_0 Instances of VLMs provide overlay/interaction suitability ratings that deviate more extreme than those provided by individual PTPs in comparison to their broader population.

To analyze *QO* and *QI,* we employ bootstrap hypothesis testing [1, 23]. For each PTP and VLM, we assess if their ratings significantly deviate from the rest of PTPs for every scenario and area (using the Mann-Whitney U test). Across all 1764 bootstrap iterations, we count the percentage of instances where the ratings of an VLM diverge more often than those of a PTP and normalize this count with the number of total comparisons, which determines the p-value [23]. For both questions, we can reject H_0 (*QO:* $p < 0.04$; *QI:* $p = 0.0$), indicating that VLMs rate scenarios not significantly *more* different to all PTPs than any individual PTP.

Analyzing the distributions of VLMs and PTPs across areas and scenarios reveals that the standard deviation of PTP responses is consistently larger than that of VLM responses (*QO:* PTPs $SD = 1.72$, VLMs $SD = 1.18$; *QI:* PTPs $SD = 1.74$, VLMs $SD = 1.11$). This can also be seen in the area ratings of the subway scenario (Figure 4, middle). Its boxbplots exemplify that medians of both conditions frequently overlap (Figure 5, 5c, 5e, 5f). For areas where they do not, PTPs often exhibit an even higher standard deviation in their ratings (Figure 5).

In addition, we explored whether VLMs and PTPs provided the same reason when an area was deemed unsuitable, i.e., when the median rating for both groups fell below 3 - 'Neutral' (0.28 of areas of both *QO* and *QI).* Specifically, we compared the fraction when the mode of responses for questions *QR-O/I* was consistent across both groups. The mode for *QR-O* was identical between PTPs and VLMs in 50% of ratings (chance would be 20%). For *QR-I,* this similarity was observed in 25% of ratings.

6.5 Discussion

Our findings suggest that SituationAdapt's reasoning module is capable of assessing situations in shared social spaces not different than experienced MR users. When evaluating both the suitability of overlays and the appropriateness of interactions, instances of VLMs did not assign more extreme ratings to situations than PTPs.

Our analysis also revealed that VLMs consistently assigned high ratings for overlay suitability to areas featuring any type of display ($MD = 5, SD = 0.72$), regardless of the context or the display's status (on or off). In comparison, PTPs' assessments of display overlay suitability varied ($MD = 4, SD = 1.62$), showing that participants

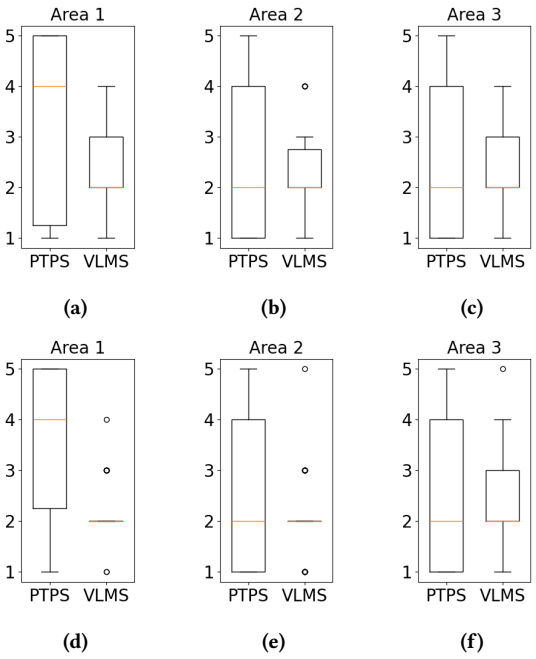


Figure 5: Boxplots of the overlay (a–c) and interaction (d–f) suitability ratings of participants (PTPs) and VLM (VLMs) for the subway scenario (Figure 4, middle). For both questions, it can be seen that the standard deviation of PTP responses is consistently larger than that of VLM responses. The boxplots further show that medians of both conditions frequently overlap (b, c, e, f). For areas where they do not, PTPs often exhibit a high standard deviation in their ratings (a).

Figure 6: Our study setup replicated a university seminar room, where the participant was sitting in the last row and another attendee was seated in the row ahead.

took contextual factors, such as whether the display was active, into account. To mitigate the influence of this bias from VLMs on our findings, we removed all areas with displays from our analysis (affecting three areas in total). Furthermore, we implemented a refinement in the context prompt provided to the VLM (added the sentence: "When a monitor displays content, overlaying a virtual element on top of it is unsuitable."). We used the new context prompt to generate results for the MR layout study and the applications.

While our statistical analysis showed that VLMs did not assign more extreme ratings than PTPs, their reasoning about the unsuitability of certain areas for UI element placement differs (with 50%- [*QR-O*] and 25% overlap [*QR-I*], respectively, by a 20% coincidence rate). It is important to mention that we did not specifically fine-tune VLMs for reasoning responses, as our system is mainly concerned about suitability ratings. Consequently, we anticipate that the reasoning alignment between VLMs and PTPs would also increase through a fine-tuning process.

7 EVALUATION OF MR LAYOUT ADAPTATION

To evaluate if our approach generates MR layouts that better adapt to situations in shared spaces, we compared it with two baseline adaptation methods. Our study thus investigated the impact of our approach on the positioning of UI elements within shared spaces, taking into account their (1) overlay in the physical environment and (2) assessing the ease of direct interaction with them.

7.1 Study design

We used a within-subject design with two variables: TASK (2 levels: *listening comprehension*, *discussion*), and METHOD TYPE (3 levels: *UserCentric*, *SurfaceAdapt*, *SituationAdapt*). For each displayed UI element, we collected participants rating for its *overlay*- and *interaction suitability*. Thus, we slightly adjusted the questions of the survey (overlay suitability: "Please rate the suitability of displaying the [UI element] where it was in this room."; interaction suitability: "Please rate how acceptable you found the direct interaction with the [UI element] given your surroundings and the people and objects in it."). Responses were recorded using a 5-point Likert scale,

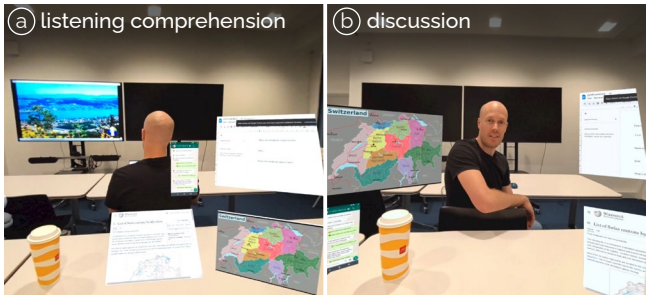


Figure 7: Participants' perspective of the adapted MR UI during the (a) *listening comprehension* and (b) *discussion* task.

with options ranging from 1 - "Unsuitable" to 5 - "Perfectly Suitable". Participants were asked to score the suitability considering the **FASH** factors of the user interface. Therefore, they were introduced to these factors at the beginning of the study. The TASK order was fixed while METHOD TYPE orders were fully counterbalanced.

Environment. Mimicking a shared social space, we ran the study in a seminar room of a university (Figure 6). The participant was seated in the last row. In the row before them the experimenter acted as another person attending the lecture. Depending on the task, they either watched the lecture or engaged with the participant. The lecture was played on a large screen before both of them. This environment included several **FASH**-relevant features, including functionality (display on/off), social acceptability (person facing towards/away), and health & safety considerations (a drink that could be spilled if occluded). These factors were the most frequently cited in our first study and are common in other real-world scenarios.

Tasks. The study involved two tasks typical of a lecture setting and purposely designed to alter the context within the seminar room. Participants first performed a *listening comprehension* task, which involved watching a geography lecture on a state of Switzerland and answering simple questions (e.g., number of inhabitants of a state) by typing the answers into a notepad widget of the MR UI (Figure 7a). In this task, participant and the experimenter were focusing on the video lecture playing on the large screen.

Subsequently, participants performed the *discussion* task. Thus, the experimenter turned around and engaged in a conversation with the participant. They asked the participant two questions about the geography of the state (i.e., highest point of elevation, number of lakes). The participant could answer the question by scrolling through a Wikipedia page or looking at the map of Switzerland. Both were provided as widgets in the MR UI (Figure 7b).

Methods. We compared *UserCentric*, *SurfaceAdapt*, and *SituationAdapt*. All conditions were implemented using AUIT [15]. To ensure a fair comparison, all conditions were made aware of the objects (TV screens, paper cup, desk), the available free spaces, and the person present in the participant's surroundings. Thus, we manually aligned the virtual and physical environments and represented objects as 3D bounding boxes within the AUIT optimization space.

UserCentric places elements in a sphere around the user using the distance, field of view, look at, occlusion terms, constant view size of AUIT. In addition, we set a physical anchor for the keyboard to align

with the desk. The weights assigned to the various factors are as follows: occlusion is weighted at 0.3, look at at 0.1, distance at 0.15, field of view at 0.3, and constant view size at 0.15. It is comparable to how virtual environments are displayed on commercial platforms such as Meta Quest or Apple Vision Pro.

SurfaceAdapt uses the same terms as *UserCentric*, and further incorporates the interaction term described in Equation 4. To prioritize placement of elements on surfaces, the interaction suitability score i_b was empirically set. The interaction frequency f_v of each virtual element was designed to fit its functionality. The weights assigned to the various factors are as follows: occlusion is weighted at 0.2, look at at 0.1, distance at 0.1, field of view at 0.2, constant view size at 0.1, and interaction suitability at 0.3. The condition serves as a representative baseline for methods aligning MR UI layouts with physical surfaces, as it has demonstrated usability benefits [9].

SituationAdapt represents our system's output. To ensure a stable environment across conditions, we also use the predefined physical environment with it. To attain ratings from our reasoning module, we captured a photo from the position of participants with a camera and manually annotate the 2D bounding boxes for each object of the defined physical environment. To simulate a realistic setting, we ran a separate VLM query for each participant and used the attained ratings in the MR UI optimization. . The training data split of the online survey was again added as context to the VLM. We used the same values for interaction frequency f_v than in *SurfaceAdapt*. The weights assigned to the various factors are as follows: occlusion is weighted at 0.2, look at at 0.05, distance at 0.1, field of view at 0.2, constant view size at 0.1, overlaying suitability at 0.15 and interaction suitability at 0.2.

7.2 Procedure

Participants started the study by completing a consent form and a demographic questionnaire. They then performed a training trial in which they familiarized themselves with the available UI elements and practiced interaction. During training, participants were introduced to the **FASH** factors and how they influence MR UI layouts in shared spaces. Afterwards, participants completed the conditions of the study, performing lecture and discussion tasks for each of the three adaptation methods (completing six trials in total). Participants completed a questionnaire after each trial. Finally, participants ranked the three adaptation methods according to preference. They completed sessions in under 30 minutes.

7.3 Participants

We recruited 12 participants (4 female, 8 male), ages 22–29 (M=26, SD=2.1) from a local university. They reported their frequency of using a VR/AR headset and using direct touch for MR interaction. For both questions, two participants mentioned using it several times a week, while eight indicated usage several times a month, and the remaining two participants reported less frequent usage.

7.4 Results

We analyzed the effect of METHOD TYPE across the different TASKS on *overlay suitability*, *interaction suitability*, and *method preference*. Due to the ordinal nature of our dependent variables, we assessed differences with a two-factor Aligned Rank Transform (ART) ANOVA. Post-hoc comparisons were then performed using the ART-C algorithm and Bonferroni correction.

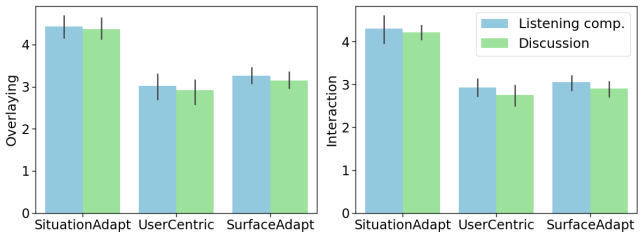


Figure 8: Mean and 95% confidence interval of participant ratings per condition for overlaying- (left) and interaction suitability (right) over all UI elements and tasks.

We found a significant effect of METHOD TYPE on *overlay suitability* across TASKS [$F_{2,66} = 67.35$, $p < .0001$]. Post-hoc tests have shown that *SituationAdapt* caused participants to perceive UI elements to be placed at more suitable locations in a shared space compared to *UserCentric* and *SurfaceAdapt* ($p < .0001$ for both, Figure 8 left). Other differences were non-significant.

Similarly, a main effect of METHOD TYPE on *interaction suitability* across TASKS was found [$F_{2,66} = 68.93$, $p < .0001$]. Post-hoc analyses revealed that participants perceived UI elements as being positioned in more interaction-friendly locations within shared social spaces when using *SituationAdapt* compared to either *UserCentric* or *SurfaceAdapt* ($p < .0001$ for both, Figure 8 right). No other significant differences were observed.

We found a main effect of METHOD TYPE on participants' preference rankings [$F_{2,66} = 143$, $p < .0001$]. Participants ranked *SituationAdapt* ($M = 1.0, SD = 0.0$) significantly higher than both *SurfaceAdapt* ($M = 2.25, SD = 0.44$) and *UserCentric* ($M = 2.75, SD = 0.44$; $p < .0001$ for both). We also found a statistically significant difference in ranking between *SurfaceAdapt* and *UserCentric* ($p < .0001$). No other significant differences were identified.

7.5 Discussion

Participants reported perceiving layouts produced by *SituationAdapt* to place UI elements at locations that are more suitable in terms of overlaying a shared space. They also perceived UI elements as being positioned in interaction-friendly locations that are suitable given the context of a shared space. Participants explained their ratings, noting that SurfaceAdapt aligns widgets with desks, making them harder to see compared to mid-air placements, and UserCentric often ignores the real-world context, frequently arranging widgets in ways that obstruct the TV or a classmate's face. In contrast, SituationAdapt avoids occluding important real-world areas, places interactive widgets on tables, and positions informational widgets in mid-air. These results suggest that *SituationAdapt* can generate MR layouts that consider the situation of the shared space surrounding the user. Moreover, the results indicate a preference for layouts generated by our method over *UserCentric* and *SurfaceAdapt*, highlighting the positive impact of adapting to the user's shared surroundings on layout preference.

8 SCENARIOS

We demonstrate SituationAdapt's ability in comprehending the context within a shared space and accordingly optimizing the placement of virtual elements across two distinct scenarios.

8.1 Discussion over lunch

We demonstrate SituationAdapt within a cafeteria setting. In this context, the user takes a break from work and is having lunch. While eating, the user watches sports videos through a virtual browser, surrounded by various virtual widgets such as sports news and messaging apps (Figure 9a). According to this initial context, all virtual elements are placed around the user, optimizing their visibility and spatial distribution. After a while, the user's colleague comes to the table, activates the laptop and starts a chat with the user. Our perception module detects the colleague and the laptop and fits the respective 3D bounding boxes (Figure 9b). The reasoning module detects that the colleague faces the user and that the laptop is turned on and provides suitability ratings. Based on these ratings, our optimization module dynamically adjusts the layout of virtual elements. All virtual elements are re-positioned away from the colleague and prevented from overlaying the laptop, ensuring the content under discussion remains unobstructed (Figure 9c). This scenario demonstrates how SituationAdapt adapts an MR UI according to the factors of 'Functionality' and 'Social acceptability'.

8.2 Preparing a meal

We demonstrate SituationAdapt in the context of a food preparation scenario. While this scenario does not feature other people, we chose it to demonstrate the usefulness of our system's adapted layouts in single-user workspaces. Within this context, the user first browses groceries, online recipes, and cooking videos in their office, where all virtual elements are placed on surfaces optimized for touch interaction. Once the user arrives in the kitchen and puts the headset on, the widgets adhere to physical surfaces according to their initial objective of facilitating interaction. As a result, the virtual elements occlude important physical objects in the kitchen, including plate and knife on the counter as well as a warning sign on the wall (Figure 9d). Our perceptual module identifies the respective objects in the environment, and extracts their 3D bounding boxes (Figure 9e). Subsequently, the reasoning module detects the best locations for placing virtual elements in the environment. Based on the ratings, our optimization module dynamically changes the layout to keep the virtual elements visible and prevent occlusion of warning sign, knife and plate (Figure 9f). This scenario exemplifies how SituationAdapt considers the factors of 'Health&Safety' and 'Functionality' when adapting MR user interfaces.

9 DISCUSSION & FUTURE WORK

We developed SituationAdapt to enable immersive interfaces to adapt to the situational context in shared spaces. In the following, we discuss limitations of our work as well as remaining open questions related to the research direction in general.

Perception of surroundings. While the implementation of our perception module is a means to an end, we still want to discuss its limitations. With RTAB-Map [2], we build on top of a traditional SLAM approach that was designed to map and navigate static environments without considering moving objects. To overcome this limitation, we manually re-initialized it with each contextual change, allowing us to retain a new 3D map per situation. Future research should use Dynamic- [58] or Semantic SLAM approaches [5] to track moving objects and people in the environment.

Our implementation of the perception model is also limited by the set of objects that YOLOv3 can recognize, which furthermore do not include large surfaces like walls. Future work should rely on objects detection methods that span a vast set of categories [63] and fuse these information with real-time segmentation approaches [59] to also gain an understanding of the surfaces in the scene.

In our current implementation, the user themselves communicates a contextual change via button press to the perception module. Future research should investigate how such a change could automatically and reliably be detected. One possible strategy could be to leverage positions and confidence values of a Semantic SLAM to discern when an object becomes relevant to the user's context.

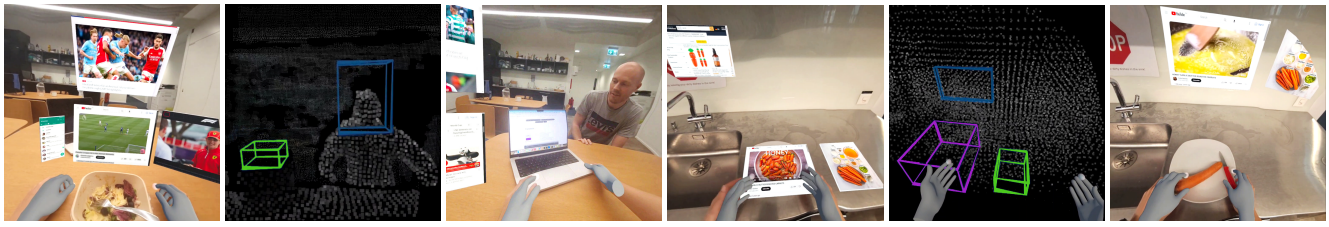
Furthermore, SDKs of future MR headsets should grant developers access to environment reconstruction and understanding features, facilitating the creation of context-aware applications without needing external hardware or redeveloping localization, mapping, and semantic understanding functionalities.

User study. We evaluated SituationAdapt in a single scenario, in which the context in a simulated lecture changed from watching a video to discussion with a classmate. However, the context of real-world shared spaces is typically more dynamic, including multiple individuals who may be strangers or friends. In addition, our user study only manipulated the shared space considering 'Health & Safety', 'Function' and 'Social acceptability' of the **FASH** factors. Future research should explore the functionality of our system in real-world shared spaces and also investigate how users perceive UI adaptations caused by all of the **FASH** factors.

VLMs for UI adaptation. In our reasoning module, VLMs utilize users' field of view as input alongside pre-designed prompts to attain human-like suitability ratings. However, users may have additional information beyond the current field of view when assessing the suitability of placing virtual elements in a shared space. For instance, users might prioritize overlaying virtual elements over strangers in public spaces while preferring to keep friends unobstructed. Inferring these relationships solely from images is unfeasible. We believe that prompting VLMs with user-specific historical data and information could help construct a context for each user and thus facilitate personalized adaptive user interfaces.

As current MR devices are designed mostly for indoor use, all scenarios in our survey were focusing on indoor environments. Initial tests in outdoor settings revealed that the VLM frequently took into account factors human evaluators considered as insignificant. Future research should investigate how VLMs comprehend shared outdoor environments and explore methods to improve their ability to accurately assess these settings. With AR glasses soon to become a mainstream consumer device, this would enable MR layout adaptation to shared outdoor spaces.

In the broader context of general HCI, we believe that our research sheds light on whether AI models are able to simulate user



(a) (b) (c) (d) (e) (f)

Figure 9: We demonstrate SituationAdapt's versatility in six use-cases: (a) the user browses sports news while having food, (b) debug output of the perception module showing the point cloud and the detected bounding boxes for the colleague (blue) and the laptop (green), (c) the virtual interface is adapted to keep the colleague and laptop unobstructed, (d) the user puts the headset on finding virtual elements to occlude the plate and the warning sign, (e) debug output of the perception module illustrating the point cloud and the bounding boxes for the sink (purple), plate (green), and warning sign (glue), (f) the virtual elements are adapted to keep the plate, sink and warning sign unobstructed.

behavior, contributing to the discourse on AI versus human reasoning. In our study, we found an interesting dichotomy in that the VLM was capable to provide ratings not different than experienced MR users, however, it struggled to provide reasoning that aligned with the rationale of these users. This aligns with findings from other studies indicating that LLMs can produce artificial responses to open-ended survey questions [53]. Future research should dive deeper into validating if AI models can simulate human participants in the context of user evaluations and further investigate the differences between human and AI reasoning.

10 CONCLUSION

We have presented SituationAdapt, an end-to-end system that considers social and environmental factors in optimizing UIs for Mixed Reality in shared spaces. SituationAdapt perceives the physical environment with real-time object detection and 3D mapping, then reasons about the suitability of placing virtual elements with a VLM, and optimizes the MR interface accordingly.

To validate our approach, we conducted an online survey where we compared VLM responses to those of experienced MR users in terms of understanding the context of shared spaces. Results suggest that the VLM judged the situations not different than participants. We then evaluated the suitability of the MR layouts generated by SituationAdapt during a lecture scenario and compared it with two baseline approaches. We found that participants rated SituationAdapt's layouts as more suitable and fitting within the situated context of the shared space.

We believe that our approach contributes an important step towards truly context- and situation-aware MR systems, enabling their adaptation to the nuances of shared social settings. We argue that this will be key to enabling MR device use in mobile settings beyond controlled home and office spaces.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Max Möbus for his assistance and support with the statistical analysis of our studies.

REFERENCES

[1] 2002. *Bootstrap Methods.* Springer New York, New York, NY, 83–96. https://doi.org/10.1007/0-387-21611-1_4
[2] 2023. RTAB-Map. http://introlab.github.io/rtabmap/
[3] Rawan Alghofaili, Michael S Solah, Haikun Huang, Yasuhito Sawahata, Marc Pomplun, and Lap-Fai Yu. 2019. Optimizing Visual Element Placement via Visual Attention Analysis. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR).* 464–473. https://doi.org/10.1109/VR.2019.8797816
[4] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7, 1 (2023), 52–62.
[5] Sean L Bowman, Nikolay Atanasov, Kostas Daniilidis, and George J Pappas. 2017. Probabilistic data association for semantic slam. In *2017 IEEE international conference on robotics and automation (ICRA).* IEEE, 1722–1729.
[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
[7] Lung-Pan Cheng, Eyal Ofek, Christian Holz, and Andrew D. Wilson. 2019. VRoamer: Generating On-The-Fly VR Experiences While Walking inside Large, Unknown Real-World Building Environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR).* 359–366. https://doi.org/10.1109/VR.2019.8798074
[8] Yifei Cheng, Yukang Yan, Xin Yi, Yuanchun Shi, and David Lindlbauer. 2021. SemanticAdapt: Optimization-Based Adaptation of Mixed Reality Layouts Leveraging Virtual-Physical Semantic Connections. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21).* Association for Computing Machinery, New York, NY, USA, 282–297. https://doi.org/10.1145/3472749.3474750
[9] Yi Fei Cheng, Christoph Gebhardt, and Christian Holz. 2023. InteractionAdapt: Interaction-driven Workspace Adaptation for Situated Virtual Reality Environments. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.* 1–14.
[10] Yi Fei Cheng, Tiffany Luong, Andreas Fender, Paul Streli, and Christian Holz. 2022. ComforTable User Interfaces: Surfaces Reduce Input Error, Time, and Exertion for Tabletop and Mid-air User Interfaces. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).* IEEE, 150–159.
[11] Hyunsung Cho, Yukang Yan, Kashyap Todi, Mark Parent, Missie Smith, Tanya R Jonker, Hrvoje Benko, and David Lindlbauer. 2024. MineXR: Mining Personalized Extended Reality Interfaces. (2024).
[12] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: visual sketching of story generation with pretrained language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts.* 1–4.
[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. (Jan. 1996).
[14] João Marcelo Evangelista Belo, Anna Maria Feit, Tiare Feuchtner, and Kaj Grønbæk. 2021. XRgonomics: Facilitating the Creation of Ergonomic 3D Interfaces. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 290, 11 pages. https://doi.org/10.1145/3411764.3445349
[15] João Marcelo Evangelista Belo, Mathias N Lystbæk, Anna Maria Feit, Ken Pfeuffer, Peter Kán, Antti Oulasvirta, and Kaj Grønbæk. 2022. Auit–the adaptive user interfaces toolkit for designing xr applications. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology.* 1–16.
[16] Andreas Fender, Philipp Herholz, Marc Alexa, and Jörg Müller. 2018. Optispace: automated placement of interactive 3D projection mapping content. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–11.

[17] Andreas Fender and Christian Holz. 2022. Causality-preserving Asynchronous Reality. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22, Article 634)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[18] Ran Gal, Lior Shapira, Eyal Ofek, and Pushmeet Kohli. 2014. FLARE: Fast layout for augmented reality applications. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 207–212. https://doi.org/10.1109/ISMAR.2014.6948429

[19] Christoph Gebhardt, Brian Hecox, Bas van Opheusden, Daniel Wigdor, James Hillis, Otmar Hilliges, and Hrvoje Benko. 2019. Learning Cooperative Personalized Policies from Gaze Data. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 197–208. https://doi.org/10.1145/3332165.3347933

[20] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1002–1019.

[21] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2016. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE transactions on visualization and computer graphics* 23, 6 (2016), 1706–1724.

[22] Jan Gugenheimer, Evgeny Stemasov, Julian Frommel, and Enrico Rukzio. 2017. Sharevr: Enabling co-located experiences for virtual reality between hmd and non-hmd users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4021–4033.

[23] Peter Hall and Susan R Wilson. 1991. Two guidelines for bootstrap hypothesis testing. *Biometrics* (1991), 757–762.

[24] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[25] Jeremy Hartmann, Christian Holz, Eyal Ofek, and Andrew D. Wilson. 2019. RealityCheck: Blending Virtual Environments with Situated Physical Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300577

[26] Anuruddha Hettiarachchi and Daniel Wigdor. 2016. Annexing reality: Enabling opportunistic use of everyday objects as tangible proxies in augmented reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1957–1967.

[27] P Jiang, J Rayan, SP Dow, and H Xia. [n. d.]. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. arXiv 2023. *arXiv preprint arXiv:2305.11473* ([n. d.]).

[28] Christoph Albert Johns, João Marcelo Evangelista Belo, Clemens Nylandsted Klokmose, and Ken Pfeuffer. 2023. Pareto Optimal Layouts for Adaptive Mixed Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 223, 7 pages. https://doi.org/10.1145/3544549.3585732

[29] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–19.

[30] Mohamed Kari, Tobias Grosse-Puppendahl, Luis Falconeri Coelho, Andreas Fender, David Bethge, Reinhard Schütte, and Christian Holz. 2021. TransforMR: Pose-Aware Object Substitution for Composing Alternate Mixed Realities. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 69–79.

[31] Mohamed Kari and Christian Holz. 2023. HandyCast: Phone-Based Bimanual Input for Virtual Reality in Mobile and Space-Constrained Settings via Pose-and-Touch Transfer. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 528, 15 pages. https://doi.org/10.1145/3544548.3580677

[32] Sagi Katz, Ayellet Tal, and Ronen Basri. 2007. Direct visibility of point sets. *ACM Transactions on Graphics* 26, 3 (July 2007), 24. https://doi.org/10.1145/1276377.1276407

[33] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health* 2, 2 (2023), e0000198.

[34] Wallace S Lages and Doug A Bowman. 2019. Walking with adaptive augmented reality workspaces: design and usage patterns. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 356–366.

[35] Jingyi Li, Ceenu George, Andrea Ngao, Kai Holländer, Stefan Mayer, and Andreas Butz. 2021. Rear-seat productivity in virtual reality: Investigating vr interaction in the confined space of a car. *Multimodal Technologies and Interaction* 5, 4 (2021), 15.

[36] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 147–160. https://doi.org/10.1145/3332165.3347945

[37] Feiyu Lu and Yan Xu. 2022. Exploring spatial UI transition mechanisms with head-worn augmented reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[38] Weizhou Luo, Anke Lehmann, Hjalmar Widengren, and Raimund Dachselt. 2022. Where Should We Put It? Layout and Placement Strategies of Documents in Augmented Reality for Collaborative Sensemaking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 627, 16 pages. https://doi.org/10.1145/3491102.3501946

[39] Tiffany Luong, Yi Fei Cheng, Max Möbus, Andreas Fender, and Christian Holz. 2023. Controllers or Bare Hands? A Controlled Evaluation of Input Techniques on Interaction Performance and Exertion in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* 29, 11 (2023), 4633–4643. https://doi.org/10.1109/TVCG.2023.3320211

[40] Lynn McAtamney and E Nigel Corlett. 1993. RULA: a survey method for the investigation of work-related upper limb disorders. *Applied ergonomics* 24, 2 (1993), 91–99.

[41] Mark McGill and Stephen Brewster. 2019. Virtual reality passenger experiences. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*. 434–441.

[42] Mark McGill, Julie Williamson, Alexander Ng, Frank Pollick, and Stephen Brewster. 2020. Challenges in passenger use of mixed reality headsets in cars and other transportation. *Virtual Reality* 24 (2020), 583–603.

[43] Daniel Medeiros, Romane Dubus, Julie Williamson, Graham Wilson, Katharina Pöhlmann, and Mark Mcgill. 2023. Surveying the Social Comfort of Body, Device, and Environment-Based Augmented Reality Interactions in Confined Passenger Spaces Using Mixed Reality Composite Videos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–25.

[44] Daniel Medeiros, Mark McGill, Alexander Ng, Robert McDermid, Nadia Pantidi, Julie Williamson, and Stephen Brewster. 2022. From shielding to avoidance: Passenger augmented reality and the layout of virtual displays for productivity in shared transit. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3640–3650.

[45] Manuel Meier, Paul Streli, Andreas Fender, and Christian Holz. 2021. TapID: Rapid Touch Interaction in Virtual Reality using Wearable Sensing. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 519–528.

[46] Roberto A. Montano Murillo, Sriram Subramanian, and Diego Martinez Plasencia. 2017. Erg-O: Ergonomic Optimization of Immersive Virtual Environments. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 759–771. https://doi.org/10.1145/3126594.3126605

[47] Aziz Niyazov, Barrett Ens, Kadek Ananta Satriadi, Nicolas Mellado, Loïc Barthe, Tim Dwyer, and Marcos Serrano. 2023. User-driven constraints for layout optimisation in augmented reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[48] Joseph O'Hagan, Julie R Williamson, Florian Mathis, Mohamed Khamis, and Mark McGill. 2023. Re-evaluating vr user awareness needs during bystander interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[49] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. 2023. Examining zero-shot vulnerability repair with large language models. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2339–2356.

[50] Xun Qian, Fengming He, Xiyun Hu, Tianyi Wang, Ananya Ipsita, and Karthik Ramani. 2022. ScalAR: Authoring Semantically Adaptive Augmented Reality Experiences in Virtual Reality. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 65, 18 pages. https://doi.org/10.1145/3491102.3517665

[51] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. http://arxiv.org/abs/1804.02767 arXiv:1804.02767 [cs].

[52] Meta Reality Labs Research. [n. d.]. SceneScript: an AI model and method to understand and describe 3D spaces. https://www.projectaria.com/scenescript/

[53] Albrecht Schmidt, Passant Elagroudy, Fiona Draxler, Frauke Kreuter, and Robin Welsch. 2024. Simulating the Human in HCD with ChatGPT: Redesigning Interaction Design with AI. *Interactions* 31, 1 (2024), 24–31.

[54] Paul Streli, Jiaxi Jiang, Juliete Rossie, and Christian Holz. 2023. Structured Light Speckle: Joint Ego-Centric Depth Estimation and Low-Latency Contact Detection via Remote Vibrometry. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3586183.3606749

[55] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[56] Priyan Vaithilingam, Tianyi Zhang, and Elena L Glassman. 2022. Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In *Chi conference on human factors in computing systems extended abstracts*. 1–7.

[57] Rafael Veras, Gaganpreet Singh, Farzin Farhadi-Niaki, Ritesh Udhani, Parth Pradeep Patekar, Wei Zhou, Pourang Irani, and Wei Li. 2021. Elbow-anchored interaction: Designing restful mid-air input. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[58] Chieh-Chih Wang, Charles Thorpe, Sebastian Thrun, Martial Hebert, and Hugh Durrant-Whyte. 2007. Simultaneous localization, mapping and moving object tracking. *The International Journal of Robotics Research* 26, 9 (2007), 889–916.

[59] Jian Wang, Chenhui Gou, Qiman Wu, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Rtformer: Efficient design for real-time semantic segmentation with transformer. *Advances in Neural Information Processing Systems* 35 (2022), 7423–7436.

[60] Julie R Williamson, Mark McGill, and Khari Outram. 2019. Planevr: Social acceptability of virtual reality for aeroplane passengers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[61] Graham Wilson, Mark McGill, Daniel Medeiros, and Stephen Brewster. 2023. A lack of restraint: Comparing virtual reality interaction techniques for constrained transport seating. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2390–2400.

[62] Jackie Yang, Christian Holz, Eyal Ofek, and Andrew D. Wilson. 2019. DreamWalker: Substituting Real-World Walking Experiences with a Virtual Reality. In *Proc. 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1093–1107. https://doi.org/10.1145/3332165.3347875

[63] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*.

A CONTEXT PROMPT OF LLM

We utilized the following prompt to establish the context for the Large Language Model (LLM). This example focuses on setting the context for overlay suitability, whereas the prompt for interaction suitability was similar.

> "You will mimic a participant of a survey in which participants had to rate the suitability of Mixed Reality layouts that overlay User Interfaces onto parts of the real world. Thus, you will rate the suitability of directly interacting with virtual UI elements that you imagine be placed on each highlighted area of an image.All virtual elements would only be visible to you, not to other people in the image. All virtual elements would not obstruct the view of other people or light. The people you can see in the image are someone else, not yourself. You will rate the suitability of each area on a score that ranges from 1 to 5 where 1 means 'unsuitable', 2 means 'somewhat unsuitable', 3 means 'neutral', 4 means 'somewhat suitable' and 5 means 'suitable'.
> You will be asked to give the primary reason for your choice of suitability. Optional reasons are: functionality, social, health & safety, aesthetics, and other. Functionality means: the UI element hinders the functionality of the physical object. Social acceptability means: looking at or interacting with the UI element would be socially inappropriate. Health & Safety means: the UI element occludes safety critical information or may lead to sanitation issues during interaction. Aesthetics means: the UI element impairs the visual appeal of the physical surroundings. Other means: your primary reason is not covered in the list above.
> To improve your ability to imitate a participant, you will be shown images they have evaluated and receive information about the median and standard deviation of their ratings for the highlighted areas of these images. Please take these ratings into account when judging new images."

The following prompt was utilized to provide the LLM with an understanding of how a group of users evaluated specific areas of a certain image (the numerical data is illustrative).

> 'Participants of a survey provided the following median responses along with standard deviations for the direct interaction suitability of the areas in this image: area 1: median 2.0, standard deviation 1.74; area 2: median 1.0, standard deviation 1.52; area 3: median 4.0, standard deviation 1.78;'