Cam-2-Cam: Exploring the Design Space of Dual-Camera Interactions for Smartphone-based Augmented Reality

Brandon Woodard
Brown University
Providence, Rhode Island, USA
brandon_woodard@brown.edu

Melvin He
Brown University
Providence, Rhode Island, USA
melvin_he@brown.edu

Mose Sakashita
Fujitsu Research
Pittsburgh, Pennsylvania, USA
msakashita@fujitsu.com

Jing Qian
New York University
Brooklyn, New York, USA
jq2267@nyu.edu

Zainab Iftikhar
Brown University
Providence, Rhode Island, USA
zainab_iftikhar@brown.edu

Joseph LaViola
University of Central Florida
Orlando, Florida, USA
jjl@cs.ucf.edu

Abstract

Off-the-shelf smartphone-based AR systems typically use a single front-facing or rear-facing camera, which restricts user interactions to a narrow field of view and small screen size, thus reducing their practicality. We present *Cam-2-Cam*, an interaction concept implemented in three smartphone-based AR applications with interactions that span both cameras. Results from our qualitative analysis conducted on 30 participants presented two major design lessons that explore the interaction space of smartphone AR while maintaining critical AR interface attributes like embodiment and immersion: (1) *Balancing Contextual Relevance and Feedback Quality* serves to outline a delicate balance between implementing familiar interactions people do in the real world and the quality of multimodal AR responses and (2) *Preventing Disorientation using Simultaneous Capture and Alternating Cameras* which details how to prevent disorientation during AR interactions using the two distinct camera techniques we implemented in the paper. Additionally, we consider observed user assumptions or natural tendencies to inform future implementations of dual-camera setups for smartphone-based AR. We envision our design lessons as an initial pioneering step toward expanding the interaction space of smartphone-based AR, potentially driving broader adoption and overcoming limitations of single-camera AR.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**; **Human computer interaction (HCI)**; **Gestural Input**; **User Experience**.

Keywords

Augmented reality, dual-camera interaction, touchless gestures, mobile AR, user engagement, multimodal feedback.

ACM Reference Format:
Brandon Woodard, Melvin He, Mose Sakashita, Jing Qian, Zainab Iftikhar, and Joseph LaViola. 2025. Cam-2-Cam: Exploring the Design Space of Dual-Camera Interactions for Smartphone-based Augmented Reality. In *ACM Symposium on Spatial User Interaction (SUI '25), November 10–11, 2025, Montreal, QC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3694907.3765919

SUI '25, Montreal, QC, Canada

ACM ISBN 979-8-4007-1259-3/25/11
https://doi.org/10.1145/3694907.3765919

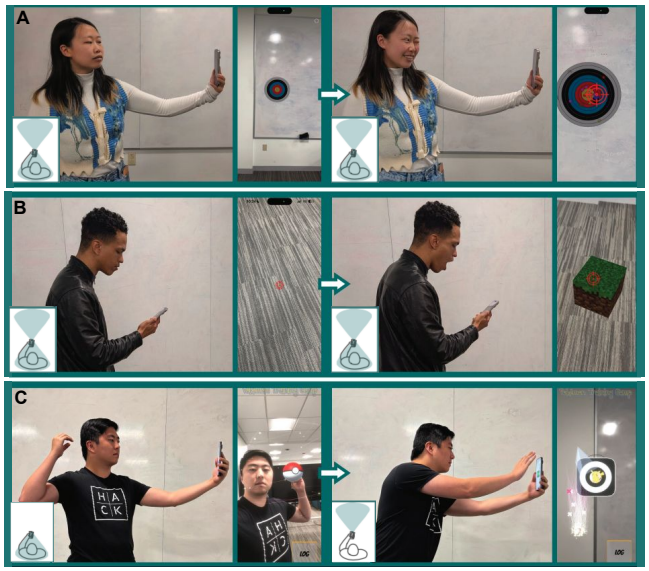

Figure 1: Cam-2-Cam combines front and rear-facing cameras to expand smartphone AR (Left), shown through three apps: A. Face TriggAR (wink), B. Mouth Craft (mouth open), C. Mirror ThrowAR (free-hand) (Right).

1 Introduction

Designing dynamic interactions for smartphone-based AR has traditionally been constrained by single-camera configurations typical in mobile devices, which either solely rely on touch-based input (e.g.,Pokemon Go [33]) or only realize the front-facing selfie camera for AR content (e.g.,Snapchat facial filters [42]) [10, 24].

Other AR form factors such as HMDs have used multiple cameras to extend the interaction space, typically allowing for more of the users' physical gestures to be captured and to expand their AR environment where users can have more ***immersive*** experiences, a common goal for AR interfaces. In contrast, smartphone AR usually

only leverages either the front-facing or the rear-facing camera and not both in tandem, which constrains the AR environment for smartphone AR [28, 36, 38, 51]. For this work, we adopt the definition of immersion presented by Agrawal et al., which states that immersion is *"A state of deep mental involvement in which the individual may experience disassociation from the awareness of the physical world due to a shift in their attentional state."* [1].

To use both cameras on the smartphone meaningfully, we delegate the front-facing camera to capture a touchless physical gesture since it already faces the user, and we render the virtual reaction in the rear camera feed, situating elements in front of the user along their line of sight. This approach builds on how humans would naturally interact with objects placed in front of us in a physical space [10, 20, 27, 36, 37]. We aim to investigate the interaction possibilities that can motivate research opportunities centered on circumventing the limited field of view (FOV) and small screen size, which can hinder the AR interaction experience on smartphones [20, 37, 51].

To this end, we present the *Cam-2-Cam* interaction concept, implemented in three dual-camera AR applications that utilize different camera input modalities (face or hand gestures) and camera configurations (alternating or simultaneous capture) to capture the breadth of the dual-camera interaction space. Each application provides a unique example of interaction types feasible with dual-camera systems: *Face TriggAR* uses a wink gesture (i.e. one-eye-closed) to simulate aiming and shooting mechanics, *Mouth Craft* enables block placement with mouth gestures, and *Mirror ThrowAR* integrates free-hand throwing detection across front and rear cameras (see figure 1). In addition to expanding on gesture-based interactions, *Cam-2-Cam* draws on the concepts of responsive multimodal feedback, incorporating haptic, visual and auditory cues to enhance users' perception of control and cohesion across dual-camera setups, informed by past work on multimodal feedback integration in AR systems [2, 3, 10, 36].

Our primary goals of this study are to construct design lessons from participants' responses that can connect the two physical spaces captured by the front and rear-facing cameras, creating a *cohesive* or unified AR interaction space. We define *cohesive* throughout this paper as a sense of connection between states or scenes, which is analogous to research centered on virtual transitions in XR and camera techniques in cinema [7, 29, 35, 39, 40, 44]. Due to the novelty of dual-camera setups for smartphone-based AR interactions, our study takes the form of an exploratory experiment where we use varying input modalities and camera techniques to cover the breadth of the space, highlight key design lessons in expanding the smartphone AR interaction space, and the significance of a larger interaction space concerning the goals of AR research.

We conducted the exploratory user study with 30 participants who partake in the *Cam-2-Cam* experience where users try all three applications in order of increasing gesture complexity to explore the implications of dual-camera interactions, identifying key themes through qualitative analysis by answering the following research questions:

- **RQ1**: How should we design the interplay between the front and rear cameras to create a cohesive interaction space?
- **RQ2**: What behaviors do users naturally adopt that are not part of the interface or experiment instructions that are helpful for the dual-camera setup?

Our qualitative analysis revealed insights into what UI techniques can extend the interaction space for smartphone AR and how the larger interaction space may lead to more user engagement, leading to wider adoption of smartphone AR. We identified three primary themes to create effective dual-camera AR interactions, which we detail below.

(1) Dual-Camera Interplay: Emphasizes the importance of creating AR interactions with real-world contexts that can translate to the form factor of dual-camera, smartphone AR.
(2) Reinforcing Dual-Camera Interactions: Virtual reactions that deliver high-quality feedback to users, creating a sense of connection between their physical space, as captured by the front-facing camera, and the AR elements displayed in the rear-camera feed.
(3) Cohesive Interaction Across Cameras: A sense of fluid interactions between cameras.

Our themes inform design lessons centered on how an expanded interaction space may make the AR experience more engaging for users and how to improve the dual-camera setup in the future. We made sure our three applications would cover the breadth of the space by implementing varying input modalities and camera techniques, including (1) camera switching, (2) simultaneous capture, (3) free-hand manipulation, and (4) facial gestures as input. The rationale for each design choice is detailed further in Section 2.3.

We make the three *Cam-2-Cam* applications available as open source: **https://bjwoods.github.io/Cam-2-Cam/**.

2 Related Work

2.1 Smartphone AR Interactions with Touchless Gestural Input

Touchless gestural input has been widely explored in smartphone-based AR to provide a user experience that is natural to users based on familiar actions they'd perform in real life. For instance, Brasier et al. presented a study to offload parts of the on-screen GUI to allow users to interact with hand gestures, allowing more screen content to be visible [10]. Qian et al. expanded on this and designed a free-hand manipulation system to manipulate AR objects using the rear-facing camera on a smartphone where they found through an iterative design process that haptic, visual, and sound cues assisted users in their AR tasks [36]. Additionally, Loorak et al. presents hand and face interactions to interact with the on-screen UI, demonstrating users' preference for this touchless interaction versus a touch-based interface when taking selfie photos [27]. The literature above illustrates the potential of touchless interactions, but they also share drawbacks they encountered when designing an AR experience for a single camera, as well as limited screen space with a narrow field of view.

2.2 Expanding the Interaction Space of Smartphones

Expanding the camera view of smartphones has been explored to expand the user interaction space and environmental awareness.

[20] introduce GlassHands, which uses reflections from sunglasses to enable interaction around unmodified mobile devices. Yeo et al. present OmniSense, a smartphone system enhanced with an omnidirectional camera to provide better around-the-body awareness [49]. These approaches expand the field of view of mobile devices, allowing for more immersive, spatially demanding interactions – these findings share common insights with literature centered on expanding the input space of head-mounted displays with multi-camera configurations [6, 40, 43, 51]. *Cam-2-Cam* similarly seeks to broaden the interaction capabilities of smartphones, but focuses on using the existing front and rear cameras to capture different types of input and output without additional hardware.

2.3 Dual-Camera Applications in Mobile Tracking and Smartphone AR

Prior work has demonstrated how smartphone cameras working synchronously can enhance real-world conditions experiences by providing additional sensing and improving environmental understanding. For instance, Babic et al. introduced SIMO, a system that uses the front camera for face and head tracking and the rear camera to estimate depth from a distant display to approximate body tracking [8]. Nagai et al. presented HandyGaze, which uses gaze and head tracking via the front camera combined with the rear camera and pre-scanned 3D model of the environment to localize users' gaze direction to identify artwork and present relevant information on a website in a mobile browser [32].

Zhao et al. used both the front-facing and rear-facing cameras to capture more accurate ambient lighting, enabling more realistic shaders and textures for AR try-on applications [50]. To the best of our knowledge, this is the only prior work that explicitly employs a dual-camera configuration for augmented reality on smartphones. However, while Zhao et al. focus on visual fidelity, the role of dual-camera setups for smartphone AR interaction techniques remains largely unexplored. Our work addresses this gap by investigating the interaction design potential of dual-camera configurations, extending the interaction space by delegating specific functions to each camera, without modifying the smartphone's form factor.

Rather than conducting premature comparisons with single-camera AR interfaces, we first establish a foundation for the nuanced parameters, such as gesture type, transition mechanics, separate or shared input and AR rendering, and underlying user values that may lead to future comparison studies. To support this, we implemented varying input modalities (face and hand gestures) and camera techniques (simultaneous capture and alternating views) to surface design insights specific to dual-camera smartphone AR.

3 Cam-2-Cam Design

Face TriggAR, *Mouth Craft*, and *Mirror ThrowAR* serve as design probes for the *Cam-2-Cam* interaction concept. Each application expands the smartphone AR input space by delegating the front-facing feed gestural input and the rear camera feed to situate AR content. We delegated the rear-facing camera to display the resulting AR elements from users' gestural actions captured by the front-facing camera because AR research and mainstream AR applications are primarily concerned with interacting with the physical world in front of the user (i.e., along the line of sight of the user). Additionally, our implemented applications incorporate audio, visual, and haptic feedback to enhance user engagement and provide immediate multimodal confirmation of actions. This design approach aligns with prior mobile AR research [28, 36, 38, 48] emphasizing the benefits of multimodal feedback and touchless gestures in improving screen occlusion caused by limited screen real estate and enhancing user perceived control in smartphone-based AR applications.

3.1 Face Gesture Task Preferences

To inform gesture-to-task mappings in *Cam-2-Cam*, we conducted a within-subjects preliminary study (N = 10) evaluating six facial gestures-wink, mouth open, smile, tongue out, head nod, and head tilt—as application input triggers across two dual-camera AR tasks: block object placement and projectile shooting.

Across the six facial gestures we evaluated (wink, mouth open, smile, tongue out, head nod, head tilt), participants most frequently selected wink (6 participants) and mouth open (3 participants) as their preferred inputs. Mouth open was chosen primarily for block placement tasks, while wink was favored for projectile shooting. Smile was selected once, and tongue out, head nod, and head tilt were not chosen by any participants. These results indicate a strong concentration of preferences around wink and mouth open, which guided our decision to implement these gestures in the *Face TriggAR* and *Mouth Craft* prototypes, respectively.

3.2 Mobile Platform Considerations

We implemented both *Face TriggAR* and *Mouth Craft* on iOS (iPhone 15) since simultaneous dual-camera capture with face tracking is currently not officially supported on Android.

To enable hand gestures for free-hand manipulation of AR content we used the *Portal-ble* free-hand AR library due to its efficacy in manipulating AR objects as shown in several studies [28, 36, 38]. The *Portal-ble* library is only supported on Android, so we implemented *Mirror ThrowAR* application on Android (using the latest Google Pixel 9) rather than iOS.

Additionally, current mobile operating systems (iOS or Android) do not allow for hand tracking combined with simultaneous capture, so we employed an alternating camera technique to capture the hand gesture with the front-facing camera and render the resulting AR content in the rear-camera feed; this approach allows us to circumvent these limitations.

3.3 Face TriggAR

Face TriggAR is developed with *Swift*, using frameworks for AR mobile scenes (*ARKit* and *RealityKit* [5]), built for iOS smartphones and later deployed on the iPhone 15. It uses world, motion, and scene tracking with a single multi-camera capture session with simultaneous front and rear-cameras for a paintball target practice simulation.

3.3.1 Scene Recognition and Target Generation: The AR scene rendered in the rear-camera view handles scene recognition, paintball shooting, ricochet with gravity projectile physics, target anchoring in world space, collision bounding boxes, and procedurally generated paint splats upon paintball collision with 3D shrapnel primitives scattering. The front camera enables eye blink motion tracking with infrared depth map (*TrueDepth* sensor on iOS) facial

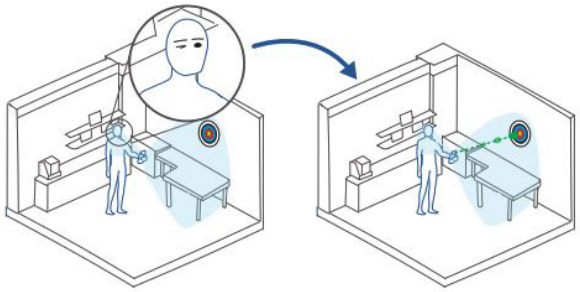

Figure 2: *Face TriggAR:* (1) Displays a zoomed-in view of the user performing the wink gesture registered by the front-facing camera. (2) Shows the resulting projectiles (virtual paintballs) being shot along the user's line of sight and visualized in the rear camera's view.

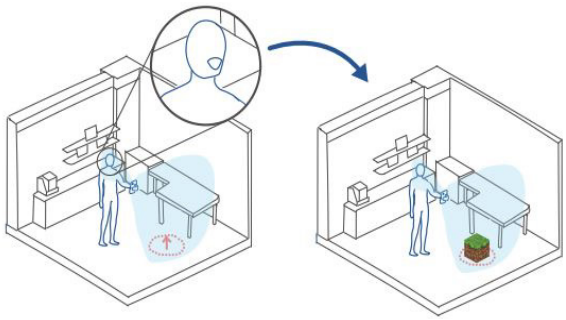

Figure 3: *Mouth Craft:* We demonstrate the (1) user performing the 'mouth open' gesture that is captured within the FOV of the front-facing camera and a detected plane is captured within the FOV of the rear-facing camera. (2) A block is placed on the detected plane and situated in the physical space captured by the rear-facing camera.

detection with which we use to continuously scan for changes in user wink input that triggers a scoping zoom with overlayed red crosshair, Western duel sound effects, and a projectile stream of paintball shots.

The system constantly scans via the rear camera for an anchor image (e.g., a QR code) to place an archery-like target while also scanning for potential eye-blink physical gestures via the front-facing camera. Users first pick up the smartphone in one hand and walk around to scan for a designated anchor image to fix the target onto a wall or other flat surface. They can also generate a fixed distance target in front of them manually without an anchor image.

Once the user creates the target, a professional-styled archery target surface (rendered as a thin cylinder) appears, see figure 2. Once the target is generated after the QR code is recognized, the target is layered flush over the anchor, sharing its surface angle.

3.3.2 Triggering Projectile Launches: When a wink is detected by the mobile device's front camera, the app zooms into the scene with a crosshair display for aiming, playing Western duel audio for dramatic effect, and continuously launching virtual projectiles every half second. Our system addresses blinks (incidental moments where both eyes are shut for less than a half second) that occur while holding a wink and adequately ignores them. The app returns to its non-shooting stagnant state when the user stops holding a wink within front camera view.

The launch direction of the small metallic blue sphere projectile is derived from the camera's forward angle, with gravity simulated. Upon target impact, haptic feedback reinforces shots, and procedurally generated 3D shards of varying colors and orientations appear on-screen momentarily, with a permanent splatter left on the target.

3.4 Mouth Craft

Mouth Craft is an AR application designed for the iPhone 15 that uses *ARKit* and *RealityKit* frameworks with *Swift* to create a touchless interaction system driven by mouth gestures for object placement. Users can build virtual 3D structures by placing Minecraft-styled [30] grass blocks into their scene environment with the front-facing camera continuously tracking facial expressions. This design allows users to act in real-time without transition delays, making the AR environment feel more immediate and responsive. When a user opens their mouth, the blocks are situated in the AR scene and rendered to the rear camera's feed.

3.4.1 Mouth Input: Users can place a block by simply opening their mouth as shown in figure 3. The system captures the user's facial expressions, detects mouth-opening gestures, and uses facial recognition algorithms with an infrared depth map to monitor the degree to which the user's mouth is open. When the system detects that the user's mouth has opened, a virtual cube is stacked into the AR scene.

3.4.2 Block Scene Placement: Once the user opens their mouth within front camera view, a block is situated in the physical space captured by the rear-facing camera. The cube's position is determined by the direction of the phone, and when the user opens their mouth, a forward-facing vector is projected along the direction of the rear camera from the center of the phone screen; then the block is placed in the physical space captured by the rear-facing camera where horizontal and vertical planes are detected by the LiDAR sensor on the iPhone. If other blocks are intersected by the vector when making the gesture, a new block will be placed on top of or flush against the previous one. Additionally, a crosshair is centered on the phone screen to assist users in placement fidelity.

When a block is successfully placed, users receive haptic feedback in the form of a slight vibration, and an auditory cue, such as a 'pop' sound, is played. This combination of tactile and auditory feedback reinforces spatial interaction. Additionally, the system enforces a brief cooldown period of half a second between consecutive block placements to ensure that gesture triggers are intentional.

3.5 Mirror ThrowAR

3.5.1 Throwing Implementation: Mirror ThrowAR builds upon an open-source smartphone AR system, Portal-ble, to enable free-hand throwing detection [36]. The system uses the Google Pixel 9's front-facing and rear-facing camera to track the user's spatial hand positions and convert them to 3D coordinates in the AR application in

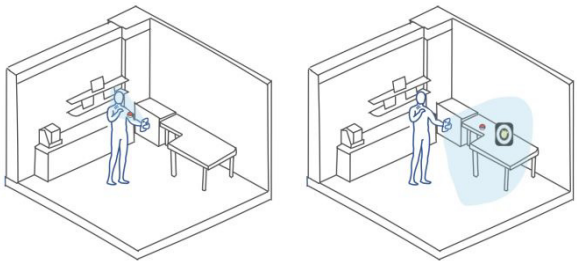

Figure 4: *Mirror ThrowAR* users hold their hand out in a pinch or ball grabbing gesture to generate a Poké Ball fixed to their hand visible on their smartphone. Then, they throw the ball toward their smartphone into the rear camera AR scene to try to hit the Pikachu target.

real time. These coordinates are used to detect overhand throwing gestures for our experiment.

We consider a three-dimensional Cartesian system, where x is the horizontal axis, y is the vertical axis relative to the virtual target generated on the wall, and z-axis is perpendicular to both x and y, pointing outwards away from the cameras making the z-axis the forward axis in respect to the target location. Given these three axes, we can define three perpendicular planes: the x-y plane, which describes the wall where the target is anchored; the y-z plane that defines the vertical space between the participant and the target; and the x-z plane, which identifies the horizontal space between the participant and the wall (see equation 1). We used ballistic motion calculations adapted from [16] :

$$\Theta_i = \arctan \frac{\left(S_i^2 + \sqrt{S_i^4 - G\left(Gx^2 + 2S_i^2\right)}\right)}{G} \tag{1}$$

where the release angle $\vec{\Theta}$ is affected by the release velocity $\vec{S}$, gravity G, The subscript i highlights the element-wise operation performed for $\vec{S}, \vec{\Theta} \in \mathbb{R}^3$ since our application operates in a 3-dimensional space.

The user performs a throw gesture toward the front-facing camera (see figures, 1 & 4), and once this action is registered, the trajectory trail continues and is displayed within the rear camera feed—creating the interaction through alternating camera views. Because the dual-camera interaction in Mirror-ThrowAR is sequential, this results in a slight latency of approximately 450 ms (0.45 seconds).

4 Formative Evaluation

4.1 Participants

We recruited participants with digital signage and email lists for on-campus graduate students at a Brown University. All participants were compensated $15 and signed a consent form before proceeding with the experiment. This study was determined to be exempt under the Institutional Review Board guidelines at Brown University.

Our recruitment goal was 30 participants as this is often sufficient to reach saturation of themes in qualitative experiments, and a higher number typically results in more nuanced insights [21, 23]. The recruited participants included 12 females and 18 males between the ages of 18 and 60 years old ($\mu = 26.97$, $\sigma = 9.41$). Participants also provided their experience with smartphone AR during our semi-structured interview and rated on a 5-point Likert scale (1 = No Experience, 5 = Very Experienced). The resulting levels of experience can be summarized as ($\mu = 1.93$, $\sigma = 0.89$).

4.2 Study Protocol

This study takes the form of a dual-camera AR experience where participants try the three developed applications (1) *Face TriggAR*, (2) *Mouth Craft*, and (3) *Mirror ThrowAR* with order increasing by task physical complexity. This study takes place within a 3 x 4 meter space with a table and couch present so users can situate AR objects on physical objects typically found indoors. Participants used *Face TriggAR* and *Mouth Craft* for the duration of two minutes and threw 10 projectiles with *Mirror ThrowAR*–values determined based on observed fatigue in preliminary experiment.

Each application session began with experimenters demonstrating the functions of the applications until participants felt comfortable trying on their own without training and starting the trial. In *Face TriggAR*, participants first scanned a QR code anchor image to generate a fixed AR target, spending the initial moments aiming at this anchored target. Afterwards, they could manually reset the target location such as on a table or coach, allowing them to explore aiming from various angles and even move behind the target to observe it from different perspectives. Using a wink gesture to aim and shoot at each target, participants had two minutes to interact with the scenes.

For the *Mouth Craft* application, participants were instructed to "build a fortress or structure" by placing virtual blocks within the AR environment using mouth gestures. They explored positioning blocks on surfaces like tables, couch, walls, and floors where a subtle popping audio feedback and a brief haptic cue followed each placement. Participants moved freely within the space to test block placement from various angles for the next two minutes.

In *Mirror ThrowAR*, participants completed a sequence of 10 Poké Ball throws into the AR scene. For each throw, participants began by holding the device in front-facing mode, capturing the throwing gesture through the front camera. As they released the Poké Ball, the application performed a sequential camera switch to the rear camera, allowing participants to view the projectile's trajectory through the AR space.Participants were instructed to aim each throw at a virtual target, receiving visual feedback in the form of motion trails and collision effects, along with accompanying haptic and audio cues at both release and impact.

4.3 Data Analysis

Experimenters conducted semi-structured interviews using a qualitative questionnaire informed by the AR Design Heuristics presented by Endsley et al. and tailored to our research questions (Table 1) [18]. The first and third authors then carried out a reflexive thematic analysis [11, 12]. Both authors first independently performed open coding, allowing codes to emerge inductively from the data while remaining attentive to patterns related to our research questions and AR design heuristics. Axial coding was then used to group related codes into higher-level categories. Codes mentioned by only one participant were typically excluded, unless they could

be meaningfully integrated into a broader category. Two subsequent meetings were held to reconcile differences, refine code boundaries, and merge overlapping categories through discussion and consensus. This process reflects a reflexive rather than reliability-focused approach to thematic analysis, emphasizing negotiated meaning and grounding themes in participants' language and experiences. The final themes and exemplar quotes are presented in Table 2.

Table 1: Survey Questions participants informed by AR Design Heuristics. Participants were also asked to suggest application scenarios.

Survey Questions	AR Design Heuristics [18]
How did your physical actions influence virtual objects?	• Form communicates function • Fit user's physical abilities
How did translation between gestures and virtual reaction feel?	• Alignment physical-virtual • Adaptation to user motion
Moments of disorientation or disruptions (selfie & rear camera)?	• Minimize distraction • Accessibility off-screen objects
How much did physical actions make virtual objects feel present?	• Fit user environment/task • Fit perceptual abilities
Which experience did you prefer and why?	• Subjective feedback

5 Results

This section presents the findings of our thematic analysis in order to outline design lessons for our dual-camera AR setup as seen in table 2.

5.1 Context-Aware Dual-Camera Interplay

We begin by detailing codes informed by users' responses, highlighting the importance of creating AR experiences with familiar concepts that can translate into a dual-camera setup. Additionally, we highlight contrasting responses in which users sometimes accept contextual misalignment when interpreting the AR experience as surreal.

5.1.1 Contextually Aligned Front-facing Camera Input with Virtual Reactions Shown in Rear Camera: Participants (N = 16, 53%) shared positive sentiments about the applications and related them to what they've experienced in the real world or video game contexts. They enjoyed *Face TriggAR*'s 'wink-to-zoom' and holding out the phone in a way that mimics shooting in the real world, as in how they perceive shooting characteristics. For example, P16 details why they enjoyed *Face TriggAR* and explains, *"I thought that squinting one eye to get a cross (crosshair) was a cool, intuitive gesture and one that I would do in a non-AR setting,"* suggesting that their gesture, captured by the front-facing camera, and visual feedback aligns with how they might perform the action in real life. Additionally, participants observed a sense of realism and associated their gestural input to actions they perform in the real world. P28 associated this realism regarding the throwing gesture in *Mirror ThrowAR* captured by the front-facing camera and shown in the rear camera as *"it's literally throwing. That's essentially, I feel that one was probably the most real, like the physical gesture."*

In contrast, for *Mouth Craft*, some participants felt it caused a disconnect. P23 describes the disconnect while contrasting it with *Face TriggAR*: *"There's some semantic connection between closing your eyes and shooting. I couldn't build such a semantic connection between opening your mouth and actually placing Minecraft cubes."*

This disconnect described during the experiment aligns with the first *AR Design Heuristics* presented by Endsley et al. [18]. The observed disconnect is essential to note as to provide a cohesive interaction experience for a dual-camera set-up, users should feel a sense of connectedness on their side of the phone (front-facing camera) to the virtual reaction shown in the rear camera. Although, a contrasting attitude towards *Mouth Craft* was described by P8 where they tell their positive experience with the mismatch of input and AR scene: *"So that felt more surreal, but I think that kind of makes it more fun because it's kind of like 'Dr. Seussy'."* In this case, the participant referenced a children's book author, Dr. Suess, known for their otherworldly stories that aren't always aligned with real-world semantics. Therefore, P8 makes a connection between a familiar artistic form of expression to contextualize their experience with their unique subjective interpretation.

Further, concepts of contextual relevance (contextually aligned front-facing camera input and virtual reactions) are seen in some responses related to the actual form factor of the smartphone itself and the types of AR 'context' the phone can provide. P21 describes the holding of the phone and closing one eye for *Face TriggAR*: *"The entire posing of it, too, is really nice because you're holding this phone far from you. You close an eye, and it's like if you're doing archery/your gun, you hold the tool out, so it's like you're role-playing this scenario."* This response focuses on the form factor in the contextual relevance of a dual-camera setup, where P21 draws a connection between how they hold the phone away from their body, close one eye for aiming, and relates this to a real-world sport, archery.

5.1.2 Alignment of Input Types and Camera Views: Several participants (N = 17, 56%) expressed attributes of contextual relevance with the alignment of cameras or FOVs and limited screen space. P2 describes the alignment of the front-facing camera to their face, complementing the context and realism of the task: *"The interaction is like a real action, and it's really helpful to do something in this way. The selfie camera can see your eyes because it's already capturing your face."* This is complemented by P20's response to question #2 about *Face TriggAR*, *"I'm looking at the target on the screen so that one translates pretty well."* These statements may demonstrate how the front-facing camera is inherently designed for capturing one's face, and closing one eye to aim assists in creating an experience that's realistic, and proper alignment of the front-facing camera may allow users to make a stronger connection to real-world contexts. P20 seemed to have felt that alignment with the virtual target they're viewing makes sense, intuitively, to how they feel they *should* be viewing the target. P24 further explains how they feel different FOVs may make more sense with different application themes: *"I felt like I had to move the camera a lot to get information for where the blocks in my scene was ... I associate a tighter FOV with, like, more precision, which has worked really well for the task."* P24 first describes difficulty with *Mouth Craft* in not being able to see all of the blocks they placed within the single FOV and how they associate the same FOV used in *Face TriggAR* with precision, thus a

shooting game made more sense with regards to the limited front-facing and rear camera FOVs. This may implicate a consideration of the contexts of motifs or actions with respect to the FOVs in a dual-camera AR setup.

5.2 Reinforcing Dual-Camera Reactions

This theme concerns the quality of multimodal feedback participants noted, which influenced their overall experience. This theme captures the sentiments of participants related to the type of AR feedback participants experienced. We identified two distinct types of feedback participants felt influenced their experience: (1) Bodily actions and responsiveness and (2) Affirming Multimodal Feedback in Dual-Camera AR.

5.2.1 Bodily Action and Responsiveness: Most participants (N = 22, 73%) felt the responsiveness of the virtual reactions gave them a greater sense of *connection*. P13 describes *Mouth Craft* as *"I feel like the fact that the as many times as you open your mouth, each time a block, would just come up, automatically...It just felt like those are kind of linked."* This response highlighted a commonality among users when it came to the responsiveness of *Mouth Craft*, and despite the misaligned semantics of the application, the responsiveness induced feelings of a 'link' or connection between the physical gesture and the responsiveness for P13, thus creating a sense of cohesiveness between the front-facing camera input and rear camera output. This sense of a cohesive interaction is tied to the responsiveness affirming their actions. Participants also described their experience with the *Cam-2-Cam* applications as feeling 'connected' to the virtual objects rendered in the rear camera's view, and they also described a sense of 'disconnect' when an application was not as responsive.

There may also be a discrepancy when an object is unexpectedly responsive. Two participants categorized in this group, P10 and P22, also shared contrasting sentiments for one of the applications. P10 describes an unexpected response when launching projectiles in *Face TriggAR*: *"In the first experiment ... it wasn't fully responsive when I kept an eye shut. There was a bit of a disconnect."* P10 shares an issue with the responsiveness that may have been due to problems with gestural recognition during their experiment. As a result, they explicitly describe, thus contrasting, the views of users who related responsiveness to a connection. This statement can highlight a contrasting side of the spectrum regarding responsiveness and creating a cohesive interaction space - balancing may be necessary for different users where reaction time across cameras is considered when designing similar dual-camera UIs.

5.2.2 Affirming Multimodal Feedback in Dual-Camera AR:. (N=9, 30%) participants highlight the potential importance of multi-modal affirmation aligned with the AR reaction. P21 describes how the haptic vibration affirmed their actions and observed reactions: *"I have feedback from this virtual weapon, or whatever you want to call it, and I think that that made it quite convincing to me that I was like, this is, real in a sense."* In this particular quote, the user underscores how the haptic feedback created a sense of realism by affirming their input gestures and virtual reactions.

Participants also emphasized the alignment of the multi-feedback and how this contributed to a cohesive experience with the dual-camera interaction. P7 describes how the mouth gesture of opening your mouth makes intuitive sense with the 'popping' noise: *"your mouth kind of goes (pop – participant does mouth popping noise). The noise did it, too. So seeing that and hearing that, and feeling you do that was all very connected. When you go, like this (Participant demonstrates the mouth gesture) and the noise goes with it, and the visuals too."* The affirming visuals and sounds contributed to the participant's overall sense of the dual-camera interaction space feeling connected. P5 echoed this sentiment but with a focus on the sound effects: *"I think for opening your mouth and placing an object, the sound effect makes it more realistic."*

5.3 Cohesive Interaction Across Cameras

Participant responses revealed how camera transitions can affect the perceived connection of their space captured by the front-facing camera and the AR elements overlaid within the rear camera view. This can lead to a fluid interaction, thus allowing more space around the phone to be usable. The following codes that may influence camera transitions uncovered in our analysis are visualizing AR physics, seamlessness, and responsiveness of alternating cameras. We define *seamlessness* as participants' sentiments related to smooth or gradual changes that make intuitive sense between their gestural input and virtual response, such as a fade-in or fade-out. *Responsiveness* in this context refers delay at which the two cameras take to alternate between each other.

5.3.1 Seamlessness of camera transitions: Uninterrupted changes between cameras or seamlessness were described by (11, 37%) participants. When asked about their feelings towards their physical gestures and virtual reactions (Q2), P14 describes an obvious *cause-and-effect* between the front-facing camera delegated to capture their gesture and rear-facing camera intended to display the AR reactions: *"In the first one (Face TriggAR), it was pretty much seamless... When I close one eye, it does that (virtual reaction).".* P14's sentiment seemed related and supported by P25's response: *"It felt, I felt like there was, like, a smooth transition for like, all of them, like the graphics you can see it slowly, like, appearing.".* P25 details their experience in a way similar to how seamlessness is described in camera transitions or AR/VR transitions on HMDs, where the emphasis is placed on the gradual change between states (e.g., action and reaction) [7, 29, 35, 39–41, 44, 47].

P11 echoes the sense of seamlessness and details specific features which may attribute themselves to a seamless experience in *Mirror ThrowAR*: *"after it landed on the target, it flipped back to me pretty smoothly, and the flipping from front to rear as I was throwing the ball was pretty smooth"*. P11's description is centered on the actions occurring in-game and the cadence of the cameras alternating. When a ball is thrown in *Mirror ThrowAR*, there is a visible camera transition where the frame is frozen until the rear camera's feed starts. Once started, the elements from the AR scene overlay onto the rear camera feed, and the projectile and trajectory become visible; lastly, once the ball lands, the user can observe its landing location for five seconds. The application switches back to the front camera feed so a user can begin throwing again.

Table 2: Themes and codes with Exemplar Quotes (heat map shading indicates prevalence of category)

Themes	Codes	Exemplar Quote
Context-Aware Dual-Camera Interplay **(N = 24, 80%)**	Contextually Aligned Front-facing Camera Input with Virtual Reactions Shown in Rear Camera **(N = 16, 53%)**	P24: "I've seen people that close one of their eyes to aim, and for some reason, I associate that with precision. That worked really well for that piece, specifically."
	Alignment of Input Types and Camera Views **(N = 17, 57%)**	P25: "The only thing I experienced was sometimes if you had your face too close or too far away, then the camera wouldn't capture ... I guess, kind of wouldn't be in the frame."
Reinforcing Dual-Camera Reactions **(N = 22, 73%)**	Bodily Actions and Responsiveness **(N = 22, 73%)**	P9: "I feel like the translation was the best for the open-mouth task (*Mouth Craft*); that was the most fluid. — I think I preferred the mouth one because it felt the most responsive."
	Affirming Multimodal Feedback in Dual-Camera AR **(N = 10, 30%)**	P5: "For opening your mouth and placing an object, the sound effects make it more realistic. The sound effects sound like a person opening their mouth."
Cohesive Interaction Across Cameras **(N = 19, 63%)**	Seamlessness of Camera Transitions **(N = 18, 60%)**	P25: "I felt like there was a smooth transition for all of them, like the graphics, you can see it slowly appearing."
	Visual affirmation of physics **(N = 9, 30%)**	P28: "I was playing with the Poké Ball; I could try to aim to where I'm casting it [the Poké Ball] off to in the camera frame."

Low (N) High (N)

In contrast, some participants also expressed disruptions or lack of seamlessness between cameras. P1 suggests addressing what they perceive as a non-seamless transition in *Mirror ThrowAR* when describing any moments of disorientation (Q3): *"I'm looking at myself, I couldn't see the target initially. If there is a separate line where one is streaming with the selfie camera and the other is streaming with the rear camera, that might be better."* P1 highlights a difficulty with the alternating camera feeds when using *Mirror ThrowAR* and suggests a 'split screen' view as seen in multiplayer video games. P1's suggestion may allow the interaction space also to be extended visually, where two AR scenes can be rendered on both camera feeds, thus increasing the visibility of AR elements and their resulting interactions across cameras.

5.3.2 Visual affirmation of physics: (N=9, 30%) participants expressed how they felt assisted or disoriented by recognizing visual feedback of the AR physics implemented in the *Cam-2-Cam* applications. P28 described creating a mental map while using *Mirror ThrowAR* between the direction of the trajectory trail and Poké Ball leaving their hand versus where the ball would land on the target: *"I tried to aim where I'm casting it (Poké Ball & trajectory trail shown in rear camera feed) into the camera frame. Which is why I think I did better. Once you figure out that one, you know ... okay, I need to throw it in the general area."*. P28 used the 'casting' direction of the ball within the boundaries of the phone screen (frame), rendering the selfie camera feed to where the projectile lands in the rear camera feed. Their description of a mental map between the two spaces captured by the front-facing and rear cameras may be interpreted as an affirmation of their logical connection.

P14 describes a similar mapping enabled by visualizing physics between their physical actions and the visible trajectory of the paintball in *Face TriggAR*: *"Moving the camera up and down would also sort of change how the ball was being shot... it was like this connection between what I was doing and what was in AR."*. P14 acknowledges that the visible projectile motion intuitively made sense between their physical actions and the AR space, thereby creating a connection between themselves and the AR scene in the rear camera feed displayed on the phone.

Not all participants were positive about the visible AR physics, and most notably, P10 shares that the AR physics they observed may not have been logical responses to their physical actions when using *Mirror-ThrowAR*. P10 states, *"The physics in the virtual environment were not related to the physics happening in the real world... a harder throw didn't mean a faster Poké Ball getting thrown at the target."*. P10 expresses an opposite effect based on how they interpreted their physical actions and the AR physics, possibly causing a disconnect between the two spaces, as highlighted when they felt the Poké Ball's speed didn't align with what they've experienced in the real world. In particular, *Mirror ThrowAR* is unique in that users start throwing with the front camera feed, where their trajectory continues into the rear camera feed. P10 describes a disconnect between when they throw into the front camera feed and the resulting projectile and trajectory once the ball is released into the rear camera feed. It's also worth noting that P10's response to Q4 centered on the connection between users' physical actions and virtual objects. This sentiment further supports how the AR physics must be meaningful between the front-facing and rear cameras, or they may cause a disconnection.

6 Discussion

6.1 Design Lessons for a Cohesive Dual-Camera Interaction Space (RQ1)

We contribute to smartphone-based AR research that leverages dual-camera setups by capturing user impressions via our questionnaire aiming to identify design lessons related to cohesive interaction

between the front-facing camera and the rear-facing camera. Our concept of a coherent dual-camera interaction space adapts the definition of cohesiveness we outline in Section 1 to focus on designing a dual-camera interface where the spaces captured by the front and rear cameras can be perceived as connected to each another.

The following design lessons detail the interface and interaction design informed by our experiment, which may allow for a cohesive interaction space for dual-camera AR interactions, expanding the limited form factor without modifying the device's hardware.

6.1.1 Design Lesson – Balancing Contextual Relevance and Feedback Quality (Themes #1 and #2): We discovered an interdependent relationship between themes #1 and #2, where participants could be more accepting of a lower context awareness of the themes and interaction of the application if the feedback quality implemented was engaging enough and vice versa. Therefore, we named this design lesson, *Balancing Contextual Relevance and Feedback Quality*.

Face TriggAR and *Mirror ThrowAR*'s strengths were centered on the familiarity of users with interactions in other real-world contexts. In contrast, *Mouth Craft*, where users could not draw a connection from the real world of placing blocks by opening their mouths, still received praise for the feedback quality people experience through sound effects, haptic vibrations, near-immediate action response, and visual elements. This multimodal feedback allowed users' actions to feel 'connected' to the blocks they were placing, thus implying a connection between the user's space and the AR space overlaid onto the rear-facing camera's feed.

Molnar et al. relate the immediacy of feedback from AR elements and the type of feedback in response to actions to flow and immersion [31] . In addition, immersion in an AR context can be defined as the presence and connection to the AR elements with which users interact [31]. The majority of positive responses for *Mouth Craft* were categorized under the theme, *Reinforcing Dual-Camera Reactions*, which implies the experience of the application can be attributed to its multimodal feedback in response to users' gestures.

P5 responses captured the lack of familiar context with *Mouth Craft* but still preferred it among the three applications due to the immediacy of feedback. P5 describes *Mouth Craft* as feeling, *"weird because these two (mouth open gesture and placing a block) make no connection in my mind"*. However, P5 follows up to explain that *Mouth Craft* was their favorite, *"because it's more responsive"* and also states, *"the sound effects sounds like a person opening their mouth"*. P5's responses about *Mouth Craft* underscore a familiar and somewhat contradicting sentiment across multiple users where they still preferred *Mouth Craft* due to the immediate feedback of the AR scene.

P15 echoes the immediacy of the feedback as benefiting the translation between their physical gestures and virtual reaction (Q2): *"It felt pretty natural and consistent — when I opened my mouth, for the most part, it would place a block."* and, *"There was a connection between what I was doing with my body and what was appearing on the screen ... I noticed it most with the mouth one (Mouth Craft) where I would do a thing, and then the tiny computer in my hand would respond to it by creating something I could see."*. What P15 describes is a sense of connectedness or embodiment to the virtual objects, which plays a critical role in the overall immersion of an AR interface [19, 40].

We suggest a design model where an application's familiarity is based on real-world context and feedback quality that can be used to balance these two attributes for future dual-camera AR setups. Our results identified a spectrum between contextual relevance and multimodal feedback, and this model can inform other smartphone AR applications with a dual-camera setup. In particular, this design model can be used to gauge whether the intended functionality of an AR application can be achieved without sacrificing immersion. The smartphone AR space needs workarounds that can bring a level of immersion closer to what is experienced by its HMD counterparts that also use multiple cameras to expand the interaction space, multimodal feedback, and gesture recognition.

6.1.2 Design Lesson – Preventing Disorientation using Simultaneous Capture and Alternating Cameras (Theme #1 & Theme #3)): Preventing disorientation using simultaneous capture and alternating cameras is focused on camera transitions and alignment that are predictable for the user. This design lesson is informed by responses falling under the themes, *Context-Aware Dual-Camera Interplay* (Theme #1) and *Cohesive Interactions Across Cameras* (Theme #3).

Cohesive Interaction Across Cameras captured sentiments from users related to their own body orientation with respect to the phone and virtual content which may indicate a separate cameras or scenes could be made to feel connected with the combination of transitions and visual affirmation of physics. These concepts, which are analogous to cinematic camera techniques, suggest that researchers may discover new overlaps between XR and cinematic effects, potentially enabling more effective action triggers and navigation in AR. A study focused on different visual transitions or camera responsiveness (e.g., simultaneous capture or alternation speed) could shed light on the Just Noticeable Difference (JND) for users between their physical actions and virtual transitions in a multi-camera smartphone AR setup, helping to quantify strategies for perceptibly connecting spaces. [20, 28, 38] Our work encourages future research to explore various methods of affirming user actions—such as self-viewing PiP GUIs or visual cues—that may help create cohesive virtual spaces in which users experience a sense of continuity between their actions and the XR content.

6.2 Considerations of User Defined Strategies for Dual-Camera AR (RQ2)

We identified strategies users developed on their own that were not part of the application interface or the experiment instructions, and these strategies provide valuable insights when designing dual-camera AR setups on smartphones. These strategies may be helpful considerations when designing smartphone AR with dual-camera AR. For example, P10 felt that head orientation was being registered with *Face TriggAR*: *"From what I remember, me moving my head also changed the direction of the screen. I think that implied a sense of control there–the head movement determined the way the thing is sprung out (shot out)Face TriggAR), as opposed to a static image that plays every time."* Similarly, P8 demonstrated to the interviewers how she tried to move her head to place objects for *Mouth Craft* in the direction she wanted to, even though she was aware she needed to move her hand holding the phone to change the direction. These responses may present a design opportunity to integrate head movements into the dual-camera interface as it may be a natural response

from users. Integrating an additional parameter like head rotation may also induce feelings of engagement as more physical input, especially natural ones, has been shown to lead to more engagement in AR research [36, 48]. Prior work has explored different theories within AR contexts, like Self-Determination Theory, to support users' agency or flexibility in how they interact with technology to provide a more natural driver for sustained engagement [9, 45].

Participant responses also revealed that users created mental maps connecting their space and gestures captured by the front-facing camera to the AR scene in the rear camera's feed. Most notably, P28, P8, and P12 describe very similar mental maps when using *Mirror ThrowAR*, where they would make minor adjustments throughout their throws and eventually began understanding where their projectile landing location would be. The strategy required them to adjust based on where their hand was within the bounds of the screen as described by P12: "So if a ball doesn't, let's say, show up, or like, appear as I expected it, I would see how my hand looks on screen to try to adjust it." Design approaches involving reducing the mental load with additional multimodal cues or a PiP view to simplify a dual-camera interface could also be considered. Such cues have suggested benefits in AR and GUIs that span spatially, allowing for more intuitive controls and enhancing feelings of embodiment [4, 25, 26].

6.3 Application Scenarios

Many participants shared application scenarios that can leverage smartphone AR with a dual-camera setup as showcased in figure 1 in our appendix demonstrating breadth of potential applications. For instance, P21 described a dual-camera AR museum application where closing either eye augments an art piece you're viewing: *"So maybe with your like, if you have two eyes open, you're seeing version A, one eye open, you're seeing version B, when I the other eye version C."* This augmentation could add more interactive experience for art exhibits where the subject viewing the artwork may become more engaged with the work. This application scenario could also lead to research centered on embodiment with art pieces.

The application scenarios in our participants' responses applied to different domains, such as art experiences, training, gaming and education. The variety of potential applications demonstrates an extensive design space smartphone AR with dual-camera setups may present. It is believed in scenario-based design, if there are a variety of use-cases for a particular system a multitude of interaction methods are likely possible [13, 14, 22, 34]. Our study is just a first step in exploring the dual-camera smartphone AR design space, but we foresee a diverse range of interaction styles and potential research venues that can leverage our qualitative data to inform future experiments. For more descriptive illustrations of these application scenarios, please see Appendix A.

7 Limitations & Future Work

This study was intended as an initial exploration of the design space for dual-camera interactions, and as such we did not include a direct comparison with conventional baselines such as touchscreen input. While our focus was on highlighting underexplored possibilities of combining front and rear-facing cameras, future work could benefit from quantitative comparisons that evaluate gesture types on dimensions such as precision, efficiency, fatigue, and social acceptability. For example, structured measures: usability scales, hardware-level benchmarks, task success rates, or performance metrics, could provide clearer insights into how dual-camera input compares to established methods or among different gestures themselves, and adopt methods described in various studies comparing touch-based to free-hand gestures [15, 17, 24].

A related consideration concerns the longer-term practicality of these techniques. Because our study was conducted in a lab setting with a focus on short-term use, factors such as fatigue, novelty effects, and social acceptability were not systematically assessed. Repeated use of facial gestures may introduce physical strain, while large hand or facial movements may not be desirable in public contexts [46]. Further, interactions that feel playful or engaging in the moment (e.g., mouth-triggered block placement) may lose appeal over time if they lack a strong semantic connection to user actions. Exploring how these interactions transition from novel experiences to everyday contexts remains an important avenue for future work.

8 Conclusion

We presented the *Cam-2-Cam* interaction concept implemented in a series of dual-camera interactive smartphone AR applications to form design lessons on how to make the smartphone AR interaction space larger while achieving the goals of AR interfaces. Our takeaways consist of the two main design lessons captured from our qualitative analysis: (1) *Balancing Contextual Relevance and Feedback Quality (Themes #1 and #2)* which details how interfaces with metaphors from the real world can be balanced with high-quality feedback to induce a sense of embodiment or immersion and (2) *Preventing Disorientation using Simultaneous Capture and Alternating Cameras (Theme #1 and Theme #3)* which outlines methods that should be considered in order to prevent disorientation using the dual-camera techniques we presented in this paper. Lastly, we present user-defined strategies that serve as design considerations for future smartphone AR applications utilizing a dual-camera setup. By conducting an exploratory study in this underexplored design space, we take a necessary first step toward uncovering new interaction possibilities. We invite future research to build upon our initial analysis and interaction concepts to create more immersive, expressive, and functional dual-camera AR experiences.

Acknowledgments

We thank our participants for their thoughtful feedback, which was vital to this research.

References

[1] Sarvesh Agrawal, Adèle Simon, Søren Bech, Klaus Bærentsen, and Søren Forchhammer. 2019. Defining immersion: Literature review and implications for research on immersive audiovisual experiences. *Journal of Audio Engineering Society* 68, 6 (2019), 404–417.

[2] Leila Alem, Franco Tecchia, and Weidong Huang. 2011. HandsOnVideo: Towards a Gesture based Mobile AR System for Remote Collaboration. In *Recent Trends of Mobile Collaborative Augmented Reality Systems*. Springer, New York, NY, USA, 135–148.

[3] John Aliprantis, Markos Konstantakis, Rozalia Nikopoulou, Phivos Mylonas, and George Caridakis. 2019. Natural Interaction in Augmented Reality Context.. In *VIPERC@IRCDL*. 50–61.

[4] Fraser Anderson, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2013. YouMove: enhancing movement training with an augmented reality mirror. In

Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (UIST '13). Association for Computing Machinery, New York, NY, USA, 311–320. doi:10.1145/2501988.2502045

[5] Apple Inc. 2024. ARKit and RealityKit. https://developer.apple.com/augmented-reality/.

[6] Rahul Arora, Rubaiat Habib Kazi, Tovi Grossman, George Fitzmaurice, and Karan Singh. 2018. SymbiosisSketch: Combining 2D & 3D Sketching for Designing Detailed 3D Objects in Situ. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3173574.3173759

[7] Jonas Auda, Sarah Faltaous, Uwe Gruenefeld, Sven Mayer, and Stefan Schneegass. 2023. The Actuality-Time Continuum: Visualizing Interactions and Transitions Taking Place in Cross-Reality Systems. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 35–40.

[8] Teo Babic, Florian Perteneder, Harald Reiterer, and Michael Haller. 2020. Simo: Interactions with distant displays by smartphones with simultaneous face and world tracking. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D. Mekler. 2023. How does HCI Understand Human Agency and Autonomy?. 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 375, 18 pages. doi:10.1145/3544548.3580651

[10] Eugenie Brasier, Emmanuel Pietriga, and Caroline Appert. 2021. AR-enhanced Widgets for Smartphone-centric Interaction. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. 1–12.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[13] John M Carrol. 1999. Five reasons for scenario-based design. In *Proceedings of the 32nd annual hawaii international conference on systems sciences. 1999. hicss-32. abstracts and cd-rom of full papers*. IEEE, 11–pp.

[14] John M Carroll. 2003. *Making use: scenario-based design of human-computer interactions*. MIT press.

[15] Yuting Cheng, Zhanwei Wu, and Ruowei Xiao. 2024. Exploring Methods to Optimize Gesture Elicitation Studies: A Systematic Literature Review. *IEEE Access* 12 (2024), 64958–64979.

[16] Peter Chudinov. 2014. Approximate analytical description of the projectile motion with a quadratic drag force. *Athens J. Nat. Formal Sci* 1 (2014), 97–106.

[17] Ze Dong, Jingjing Zhang, Xiaoliang Bai, Adrian Clark, Robert W Lindeman, Weiping He, and Thammathip Piumsomboon. 2022. Touch-move-release: Studies of surface and motion gestures for mobile augmented reality. *Frontiers in Virtual Reality* 3 (2022), 927258.

[18] Tristan C Endsley, Kelly A Sprehn, Ryan M Brill, Kimberly J Ryan, Emily C Vincent, and James M Martin. 2017. Augmented reality design heuristics: Designing for dynamic interactions. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 61. Sage Publications Sage CA: Los Angeles, CA, 2100–2104.

[19] Adélaïde Genay, Anatole Lécuyer, and Martin Hachet. 2021. Being an avatar "for real": a survey on virtual embodiment in augmented reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (2021), 5071–5090.

[20] Jens Grubert, Eyal Ofek, Michel Pahud, Matthias Kranz, and Dieter Schmalstieg. 2016. Glasshands: Interaction around unmodified mobile devices using sunglasses. In *Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces*. 215–224.

[21] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82.

[22] Zainab Iftikhar, Qutaiba Rohan ul Haq, Osama Younus, Taha Sardar, Hammad Arif, Mobin Javed, and Suleman Shahid. 2021. Designing parental monitoring and control technology: A systematic review. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part IV 18*. Springer, 676–700.

[23] Zainab Iftikhar, Yumeng Ma, and Jeff Huang. 2023. "Together but not together": Evaluating Typing Indicators for Interaction-Rich Communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.

[24] Minseok Kim and Jae Yeol Lee. 2016. Touch and hand gesture-based interactions for directly manipulating 3D virtual objects in mobile augmented reality. *Multimedia Tools and Applications* 75 (2016), 16529–16550.

[25] Marc Erich Latoschik, Jean-Luc Lugrin, and Daniel Roth. 2016. FakeMi: A fake mirror system for avatar embodiment studies. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. 73–76.

[26] Nianlong Li, Zhengquan Zhang, Can Liu, Zengyao Yang, Yinan Fu, Feng Tian, Teng Han, and Mingming Fan. 2021. Vmirror: Enhancing the interaction with occluded or distant objects in vr with virtual mirrors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[27] Mona Hosseinkhani Loorak, Wei Zhou, Ha Trinh, Jian Zhao, and Wei Li. 2019. Hand-over-face input sensing for interaction with smartphones through the built-in camera. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[28] Jiaju Ma, Jing Qian, Tongyu Zhou, and Jeff Huang. 2023. FocalPoint: Adaptive Direct Manipulation for Selecting Small 3D Virtual Objects. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–26.

[29] Elisabeth Mayer, Francesco Chiossi, and Sven Mayer. 2024. Crossing Mixed Realities: A Review for Transitional Interfaces Design. *Proceedings of Mensch und Computer 2024* (2024), 629–634.

[30] Mojang Studios. 2024. Minecraft. https://www.minecraft.net/.

[31] György Molnár and Zoltán Szűts. 2019. Augmented reality, games and art: immersion and flow. *Augmented Reality Games I: Understanding the Pokémon GO Phenomenon* (2019), 61–67.

[32] Takahiro Nagai, Kazuyuki Fujita, Kazuki Takashima, and Yoshifumi Kitamura. 2022. HandyGaze: A Gaze Tracking Technique for Room-Scale Environments using a Single Smartphone. *Proceedings of the ACM on Human-Computer Interaction* 6, ISS (2022), 143–160.

[33] Niantic, Inc. 2024. Pokémon GO. https://pokemongolive.com/.

[34] Jamie O'Hare, Elies Dekoninck, Mendy Mombeshora, Philippe Martens, Niccolò Becattini, and Jean-Francois Boujut. 2020. Defining requirements for an Augmented Reality system to overcome the challenges of creating and using design representations in co-design sessions. *CoDesign* 16, 2 (2020), 111–134.

[35] Fabian Pointecker, Judith Friedl, Daniel Schwajda, Hans-Christian Jetter, and Christoph Anthes. 2022. Bridging the gap across realities: Visual transitions between virtual and augmented reality. In *2022 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 827–836.

[36] Jing Qian, Jiaju Ma, Xiangyu Li, Benjamin Attal, Haoming Lai, James Tompkin, John F Hughes, and Jeff Huang. 2019. Portal-ble: Intuitive free-hand manipulation in unbounded smartphone-based augmented reality. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 133–145.

[37] Jing Qian, David A. Shamma, Daniel Avrahami, and Jacob Biehl. 2020. Modality and Depth in Touchless Smartphone Augmented Reality Interactions. In *Proceedings of the 2020 ACM International Conference on Interactive Media Experiences* (Cornella, Barcelona, Spain) *(IMX '20)*. Association for Computing Machinery, New York, NY, USA, 74–81. doi:10.1145/3391614.3393648

[38] Jing Qian, Meredith Young-Ng, Xiangyu Li, Angel Cheung, Fumeng Yang, and Jeff Huang. 2020. Portalware: A Smartphone-Wearable Dual-Display System for Expanding the Free-Hand Interaction Region in Augmented Reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3334480.3383079

[39] Joan Sol Roo and Martin Hachet. 2017. One reality: Augmenting how the physical world is experienced by combining multiple mixed reality modalities. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*. 787–795.

[40] Mose Sakashita, Hyunju Kim, Brandon Woodard, Ruidong Zhang, and François Guimbretière. 2023. VRoxy: Wide-Area Collaboration From an Office Using a VR-Driven Robotic Proxy. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[41] Kaili Shan, Tiger Sun, Jarrod Tart, Brandon Woodard, Irene Humer, and Christian Eckhardt. 2024. Work-in-Progress—Virtual Learning Laboratories for High School Chemistry Lab: An Immersive Learning User Study. *Immersive Learning Research-Academic* (2024), 61–71.

[42] Snap, Inc. 2024. Snapchat AR Filters. https://whatis.snapchat.com/.

[43] Hemant Bhaskar Surale, Aakar Gupta, Mark Hancock, and Daniel Vogel. 2019. Tabletinvr: Exploring the design space for using a multi-touch tablet in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] Chiao-I Tseng and Chiao-I Tseng. 2013. *Cohesion in film*. Springer.

[45] April Tyack and Peta Wyeth. 2017. Exploring relatedness in single-player video game play. In *Proceedings of the 29th Australian conference on computer-human interaction*. 422–427.

[46] Shuning Wang, Linghui Zhong, Yongjian Fu, Lili Chen, Ju Ren, and Yaoxue Zhang. 2024. UFace: Your Smartphone Can" Hear" Your Facial Expression! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–27.

[47] Brandon Woodard, Margarita Geleta, Jr. LaViola, Joseph J., Andrea Fanelli, and Rhonda Wilson. 2025. AudioMiXR: Spatial Audio Object Manipulation with 6DoF for Sound Design in Augmented Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3, Article 140 (Sept. 2025), 41 pages. doi:10.1145/3749478

[48] Qinyang Wu and Chen Li. 2024. Enable Natural User Interactions in Handheld Mobile Augmented Reality through Image Computing. In *Proceedings of the 2024 ACM Symposium on Spatial User Interaction*. 1–2.

[49] Hui-Shyong Yeo, Erwin Wu, Daehwa Kim, Juyoung Lee, Hyung-il Kim, Seo Young Oh, Luna Takagi, Woontack Woo, Hideki Koike, and Aaron John Quigley. 2023.

OmniSense: Exploring Novel Input Sensing and Interaction Techniques on Mobile Device with an Omni-Directional Camera. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[50] Yiqin Zhao, Sean Fanello, and Tian Guo. 2023. Multi-camera lighting estimation for photorealistic front-facing mobile augmented reality. In *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*. 68–73.

[51] Fengyuan Zhu and Tovi Grossman. 2020. Bishare: Exploring bidirectional interactions between smartphones and head-mounted augmented reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

Appendix

A Application Scenarios

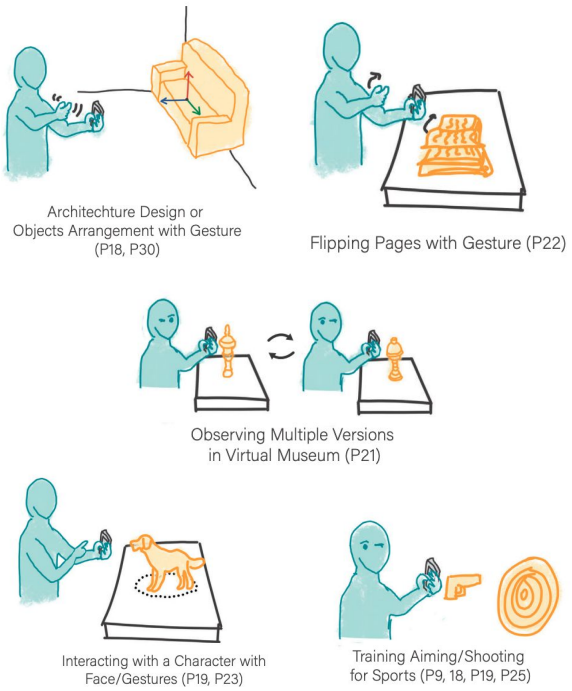

Figure 5: After trying out *Cam-2-Cam* interactions, participants shared application scenarios that they thought could greatly leverage a dual-camera setup to incorporate physical gestures.

Figure 5 illustrates the five scenarios. Together with our qualitative findings, these sketches indicate that dual-camera AR enables novel interaction methods across art, education, gaming, training, and design, motivating further empirical studies.

(1) Architecture & Object Arrangement (P18, P30): Using hand-to-camera gestures to position and manipulate virtual furniture or building elements in real space.
(2) Gesture-Based Page Flipping (P22): Winking or other face movements to turn pages in e-books, social media feeds, or document viewers.
(3) Virtual Museum Augmentation (P21): Closing one eye to switch between different versions or layers of an artwork, deepening engagement and exploring embodiment with art pieces.
(4) Character Interaction via Facial Gestures (P19, P23): Animating or commanding virtual characters or pets through smiles, winks, or other facial inputs.
(5) Sports Training & Aiming (P9, P18, P19, P25): Using precise eye- or face-based gestures to aim virtual projectiles and launch virtual projectiles.