Perspective-Aligned AR Mirror with Under-Display Camera: Supplementary Technical Report

JIAN WANG*, SIZHUO MA, KARL BAYER, YI ZHANG, PEIHAO WANG, and BING ZHOU, Snap Inc., USA
SHREE K. NAYAR, Snap Inc. and Columbia University, USA
GURUNANDAN KRISHNAN, Snap Inc., USA

In this report, we document important details that are left out from the main paper due to the page limit. **Please refer to the supplementary video to see our AR mirror in action.**

1 CALIBRATION DETAILS

Wire effect. To calibrate the wire effect, we place a shadowless LED panel in front of the display, which is composed of an array of LED lights covered by a diffuser. Since the radiance emitted by the LED panel is approximately uniform, the irradiance at the pixels is independently of the distance and orientation of the LED panel relative to the camera, and only depends on the spatial-varying irradiance modulation due to the OLED pixel structure. We capture 300 images continuously and take the average as our calibrated pattern. We also build a metal frame and rigidly attach the LED panel onto it such that the frame can sit on the display and keep stable during the calibration.

Backscatter. Since the camera is placed at eye height, the portion of display that occludes the camera usually contains the users' faces. Therefore, we chose a publicly-available face dataset (FFHQ [Karras et al. 2019]), scaled and displayed the face crops on the display to approximate the real backscatter distribution. Since FFHQ contains high-quality face images, mostly professionally captured, it does not include many overly bright or even saturated images. To enhance the network's capability of removing strong backscatter, we increase the overall intensity of 1/3 of the face images through scaling and gamma mapping:

$$I' = (a \cdot I^{\gamma} + b)^{1/\gamma}, \tag{1}$$

where $\gamma = 2.2$, $a = 2$, $b = 0.3$. Fig. 1 shows the effect of backscatter balancing. We plot the histogram for pixel intensities from all images. Intensities of the original FFHQ images center at around 0.5. After boosting the intensities of 1/3 of the images, there are clearly more pixels with high intensities.

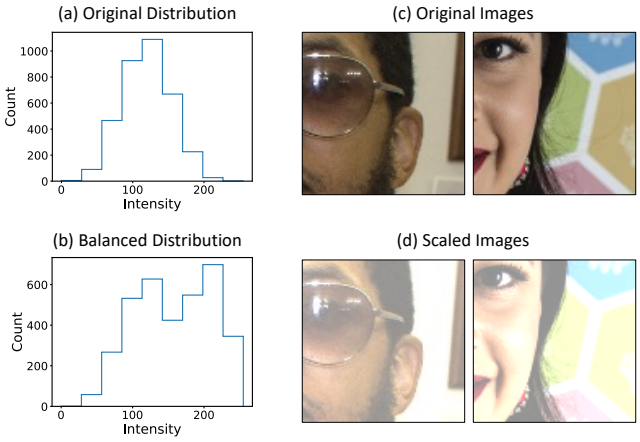

Fig. 1. Backscatter balancing. (a) Intensity distribution of the original FFHQ images. Intensity centers at around 0.5. (b) After scaling 1/3 of the images, there are more pixels with high intensities. (c)(d) Examples of original and scaled images.

2 CAMERA FRAMING DESIGN

This section outlines the camera framing design aimed at optimizing user experience, focusing on camera selection, placement, and image post-processing, including undistortion and cropping. Our design principles are: 1) Maintaining eye contact: the user's eyes should align with their image on the display when looking straight ahead, creating a sense of equal physical size and enhancing both mirroring and telepresence experiences; 2) Full body capture: the camera should capture the user's full body at 1080P resolution for applications like virtual try-on and remote training.

User distance. The design assumes users will interact with the device from about 5 feet away, a distance found to be ideal for large-format displays and interaction.

Camera height. To maintain eye contact, the camera must be positioned at the user's eye level. This placement ensures the captured image appears to make eye contact when the user looks straight into the camera. Although users have varying heights, the perception of gaze has some tolerance [Cline 1967; Gibson and Pick 1963], and users typically adjust their position to achieve eye alignment.

Camera choice. Selecting the appropriate camera and lens is crucial. It is worth noting that if an upright camera aligned horizontally with the user's eyes is used, the user's body occupies only about half of the field of view (FOV), as shown in Fig. 3 in the main paper. This setup requires a short focal length and significant cropping to achieve a 1080P resolution, necessitating a high-resolution sensor (*e.g.*, 4K). However, by tilting the camera downward, a smaller FOV can capture the full body at approximately 1080P resolution. This configuration meets our design goals more efficiently. We tested two specific combinations of sensors and lenses: 1) An 8MP sensor (Basler ace2 a2A3840-45ucBAS) with a 4mm lens (Edmund Optics 33-300), and 2) A 3MP sensor (Basler ace acA2040-120uc) with a 6mm

*Shree served as the direction lead, Gurunandan as the project lead, and Jian as the tech lead and IC (individual contributor). Sizhuo and Yi contributed equally overall. Yi and Peihao contributed equally to the image restoration experiments. Jian is the corresponding author.

Authors' addresses: Jian Wang, jwang4@snapchat.com; Sizhuo Ma, sma@snapchat.com; Karl Bayer, karlsbayer@gmail.com; Yi Zhang, zhangyi3.link@gmail.com; Peihao Wang, peihaowang@utexas.edu; Bing Zhou, bzhou@snapchat.com, Snap Inc., 229 W. 43rd St 6th Floor, New York, NY, 10036, USA; Shree K. Nayar, nayar@cs.columbia.edu, Snap Inc. and Columbia University, New York, USA; Gurunandan Krishnan, guru@gurukrishnan.com, Snap Inc., New York, USA.

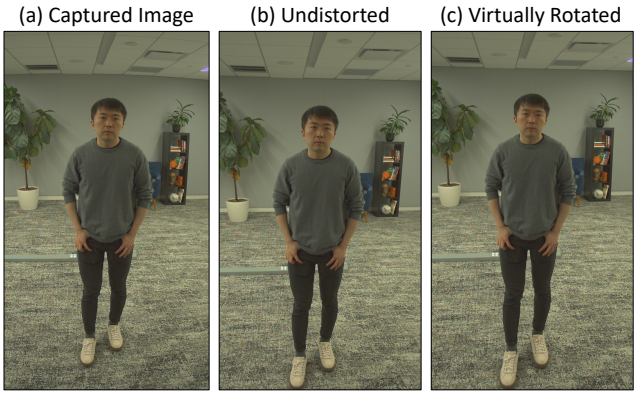

Fig. 2. Post-processing. (a) Image captured by a tilted camera distorts the body shape of the subject. (b) Lens undistortion corrects warped lines in the scenes. (c) Virtually rotating the scene via homography can correct the shortening of legs, giving a more faithful presentation of the subject.

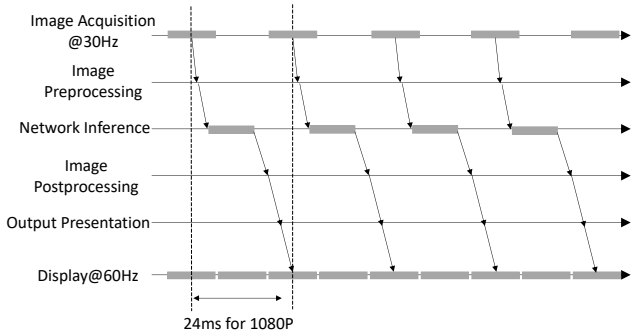

Fig. 3. Data flow and latency of our image processing pipeline.

lens (Edmund Optics 33-301). We chose the second combination for its superior overall image quality.

Post-processing for framing. To present the user correctly, we calibrate the camera's intrinsic and distortion parameters and apply undistortion to the captured image. A tilted camera can make a person's legs appear shorter due to perspective projection, so we use homography to *virtually* rotate the image. Fig. 2 shows the undistorted and rotated result, correcting the user's body shape. By adjusting the camera tilt and crop region in the final rotated image, we ensure the eye position in the displayed image aligns with the user's true eye level for an average height 5'6". Users of different heights can adjust their distance to the display, achieving an approximate eye level match.

3 AR MIRROR SYSTEM

3.1 Image Processing Pipeline

To optimize image quality, we process the pipeline at 1080P (FHD) resolution cropped from the raw image. Real-time operation at FHD resolution has two major requirements: 1) processing time <33ms to enable real-time experience, and 2) minimal lag to ensure interactive experience, both necessitating substantial computational power.

Fig. 4. Lens reflection. Without the black background, the reflection of the camera itself is visible in the captured image.

We implement a highly optimized image processing pipeline on multiple GPUs to enable real-time computation at FHD resolution. The overall frame processing pipeline, illustrated in Fig. 3, can be summarized as follows:

- *Image acquisition.* The pipeline continuously captures raw 12-bit Bayer images from a Basler camera at 30Hz via USB3 and transfers them to the GPU for parallel processing.
- *Image preprocessing.* Preprocessing involves demodulation to remove wire artifacts, demosaicing, and image tile extraction using CUDA-accelerated OpenCV. The image is demodulated with a wire pattern, demosaiced to RGB, and divided into two slightly overlapping 1152×1152 tiles for parallel processing on two GPUs.
- *Network inference.* Using TensorRT, the extracted tiles are processed by the image restoration network in parallel on two GPUs, with results copied to the primary GPU for post-processing.
- *Image postprocessing.* CUDA-accelerated OpenCV is used for tile stitching into a 2016×1152 image, followed by undistortion and virtual rotation to 1920×1080. Color adjustment is done via HSL transformations, and image enhancement includes smart sharpening and temporal post-processing using the past two frames.
- *Output presentation.* The processed image is displayed in fullscreen mode via OpenGL, saved, or written to shared memory, with options to feed into downstream applications like AR filters and video conferencing.

We use two Nvidia RTX4090 GPUs, each processing half the image with some overlap. A third RTX4090 GPU handles augmented reality (AR) effects, including face/body tracking and virtual try-on filters, as well as rendering. GPU usage is around 75% for the first two GPUs and 30% for the third. Most processing is done on the GPUs using CUDA for parallel computing. The total image processing latency from capture to display is 24 ms, showcasing the system's real-time capabilities.

3.2 Mechanical Design

Our system uses an LG-55EW5G-V transparent OLED display, a Basler acA2040-120uc camera, and an Edmund Optics 6mm/F1.85 lens. We designed and assembled a frame and facade to securely mount these components. Key design considerations include: mechanical rigidity, functional configuration and user experience. The system is designed to be portable, aesthetically pleasing, to allow configurable camera pose, and to conceal the camera as much as possible.

Mechanical rigidity. Maintaining camera/display alignment is crucial to prevent calibration from drifting and maintain image quality. We developed a rigid frame using one-inch T-slot aluminum extrusion, steel hardware, CNC-machined camera mounts, and precise opto-mechanical components. The design minimizes mechanical linkages to avoid mechanical creep during transportation.

Functional configuration. To make the camera inconspicuous, we placed a matte-black background 5mm behind the display, with a slot for the adjustable camera mount. Black felt around the lens blocks light from passing through the background. This setup hides the camera when the display is on with adequate lighting. Since light from the back is blocked, it also prevents the camera from capturing its own reflection, as shown in Fig. 4. The camera position can be adjusted vertically, in distance from the display, and in angle using manual opto-mechanical stages and precision-machined aluminum brackets. Horizontal position is fixed to the display's center.

User experience. We added a powder-coated facade to cover the sub-frame, allowing for logos or identifiers on the AR Mirror. An access door in the back facilitates camera adjustments and cable access. The facade overlaps the display by about an inch (diagonally), using a foam gasket to prevent light leaks. The display background blocks most light coming from the back the display, with the facade providing secondary protection against light pollution.

4 USER STUDY

User study design. We recruited 24 participants, ensuring a diverse range of users with various levels of technical background and experience. A $25 gift card was provided to each participant as a token of appreciation for their involvement. Participants were kept unaware of the specific details regarding the camera systems under evaluation. The study focused on evaluating the proposed Under Display Camera (UDC) system in two applications: AR mirror and video conferencing. For each application, participants interacted with two versions of AR mirrors—one equipped with the UDC system, and the other with an identical camera positioned beside the screen, named Side Camera Display (SCD). To minimize order effects, participants experienced both systems in a counterbalanced order, which were referred to as "Test A" and "Test B". Participants provided feedback through 1) Likert-style [Likert 1932] questions such as "I felt more video lag in Test A than in Test B.", and 2) open-response questions such as "Which experience (Test A or Test B), did you prefer, and why?". **See the attached screenshots in Fig. 5, 6, and 7 for detailed questions. Every participant signed a legally-reviewed consent form prior to the study.**

Quantitative results. In the evaluation of image quality, participants were prompted with specific questions, including assessments of perceived superiority in image quality, clarity of the AR mirror video, incidence of glitches, video smoothness, visibility of pixel noise on the screen, and the perception of video lag. The presented results are depicted in Fig. 10(a) in the main paper. Notably, our UDC system exhibits superior or comparable performance over the SCD across various metrics (around 3–"Neutral" for both systems), which proves that with our processing pipeline, putting the camera behind the display does not compromise perceptual visual quality. One explanation that the scores for UDC are even higher than those for SCD is that UDC gives a better overall experience, which introduces a bias when judging the image quality as well.

For the AR mirror comparison, emphasis was placed on aspects related to user comfort and engagement. Participants responded to questions regarding their comfort level and ease, the mirror's resemblance to a real mirror, the directness of eye contact, increased engagement, and the natural feel of selfies taken in the mirror. Results are presented in Fig. 10(b) in the main paper, revealing a substantial user preference in favor of our UDC system over the SCD system across all metrics. Specifically, the UDC was scored over 4.0 in almost all metrics, demonstrating the importance of user perspective and eye contact on overall user experience.

For teleconferencing, the participants were asked questions such as feeling more present in the video conference, ease of maintaining eye contact, comfort level with the chat interface, overall enjoyment of the conversation, ease of communication, naturalness in conversation, increased focus, and a sense of closeness to the person being communicated with. Results in Fig. 10(c) in the main paper reveal a clear superiority of our UDC over the SCD system across all evaluated metrics. This significant performance difference proves that the improved perspective enabled by our UDC design benefits not only AR mirror but also teleconferencing, and potentially other applications that require correct perspective and eye contact.

Qualitative feedback collection. Participants were given open-ended questions, including preferences between experiences and reasons behind their choices, details about their interaction with the mirror, aspects that felt "natural" to them, and instances that felt awkward. Representative comments from all participants are cited to encapsulate the diverse perspectives and insights gathered during the qualitative feedback collection process. In the open-ended responses, note that we've replaced user phrasing for the randomly ordered "Tests A/B" back to "UDC" and "SCD" on a per-user basis. Responses are listed below:

> "Personally preferred [UDC], because the angle fully represents my true height and true body shape. "
>
> "I preferred [UDC] by a mile. It felt much more realistic and because I wasn't as distracted by the lack of eye contact, it was easier for me to engage with the lenses themselves. I also found it to be more natural in terms of taking pictures - because I could look at my phone camera in the mirror and the position was oriented straight, like a regular mirror."
>
> "I preferred [UDC] immensely because it felt like I was in a real life fitting room. I didn't have to guess where the camera was and I could be more playful with the entire experience. "

"[UDC] felt more realistic and more like a real mirror. I felt like I could actually see my actions. "

"[UDC] because it is a better representation of what a mirror is expected to be."

"[UDC] felt more natural, [SCD] was from a non-frontal angle and felt slightly awkward."

"[UDC] used a camera that faced me directly and felt more like a mirror. [SCD] used an off-axis camera that felt more like a photographer"

"To me, the main difference between [UDC] and [SCD] was the position of the camera. In [SCD], the camera was off to the left, so when I would look directly at myself in the mirror, I wasn't making eye contact with myself, which was distracting. In [UDC], the camera position was behind the mirror itself so I was making eye contact with myself while staring directly into the mirror, which felt more natural. The difference was extremely noticeable. "

Limitation feedback. We also got very valuable feedback on the limitations such as: "It felt most natural standing around 4 feet away. It felt like a real mirror as I was the size I expected to be. Getting too close to the mirror felt awkward as the picture felt bigger than I would expect on a mirror. Moving around felt pretty natural and as expected for a mirror.", "I don't think it's that important for me. I'm already used to not seeing myself look directly at me because of taking selfies with a phone camera. But without any lenses on, it felt more like a real mirror when I was making eye contact with my self.", and "Cool, but can you use a Mac mini to handle the computing stuff, just like with the other AR mirror?"

REFERENCES

Marvin G Cline. 1967. The perception of where a person is looking. *The American journal of psychology* 80, 1 (1967), 41–50.

James J Gibson and Anne D Pick. 1963. Perception of another person's looking behavior. *The American journal of psychology* 76, 3 (1963), 386–394.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

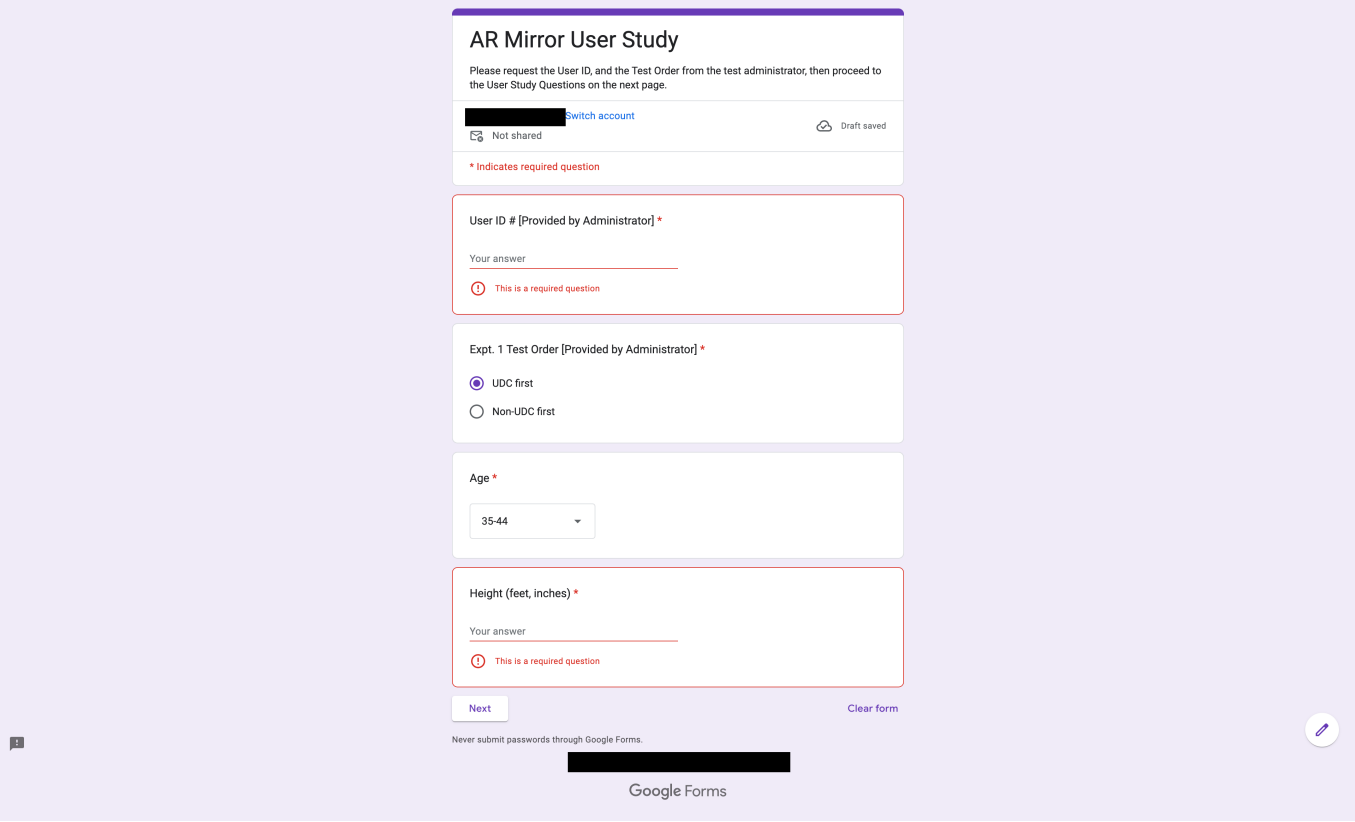

Fig. 5. User study screenshot (Page 1).

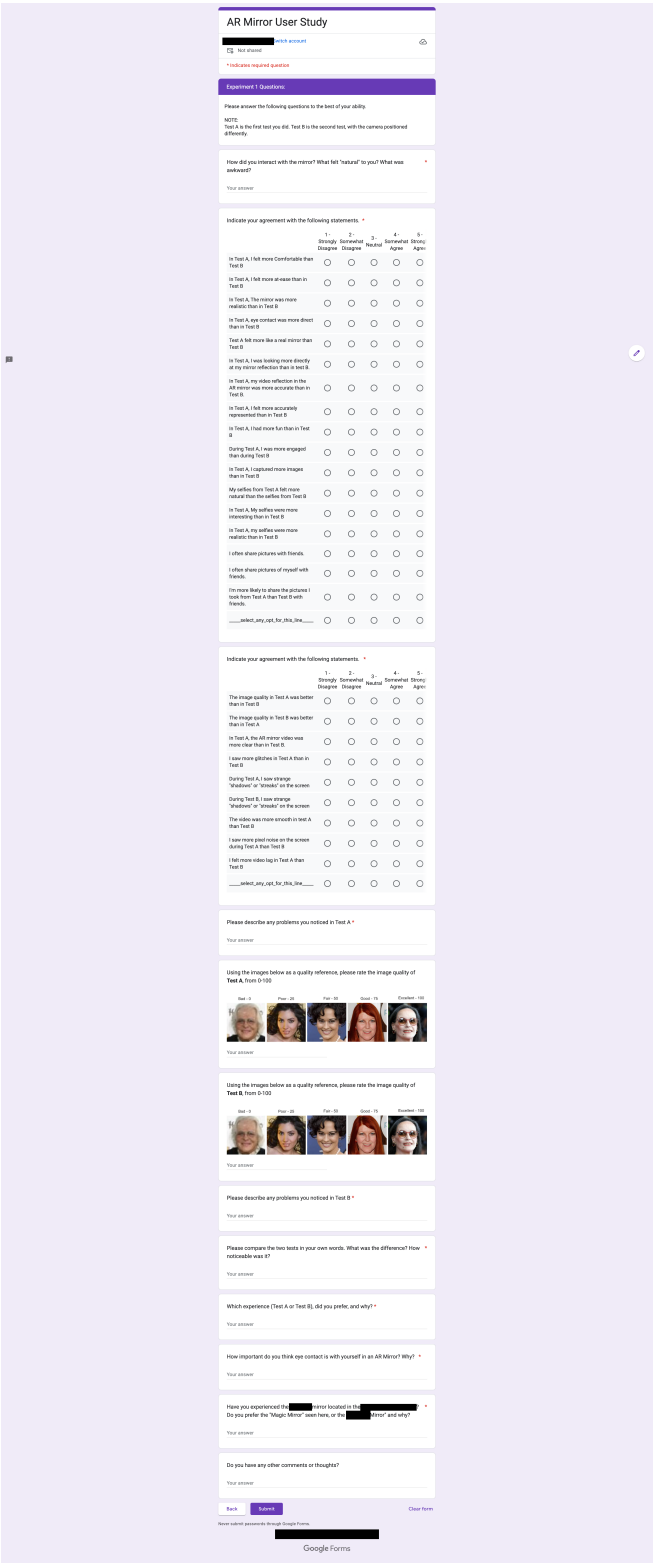

Fig. 6. User study screenshot (Page 2).

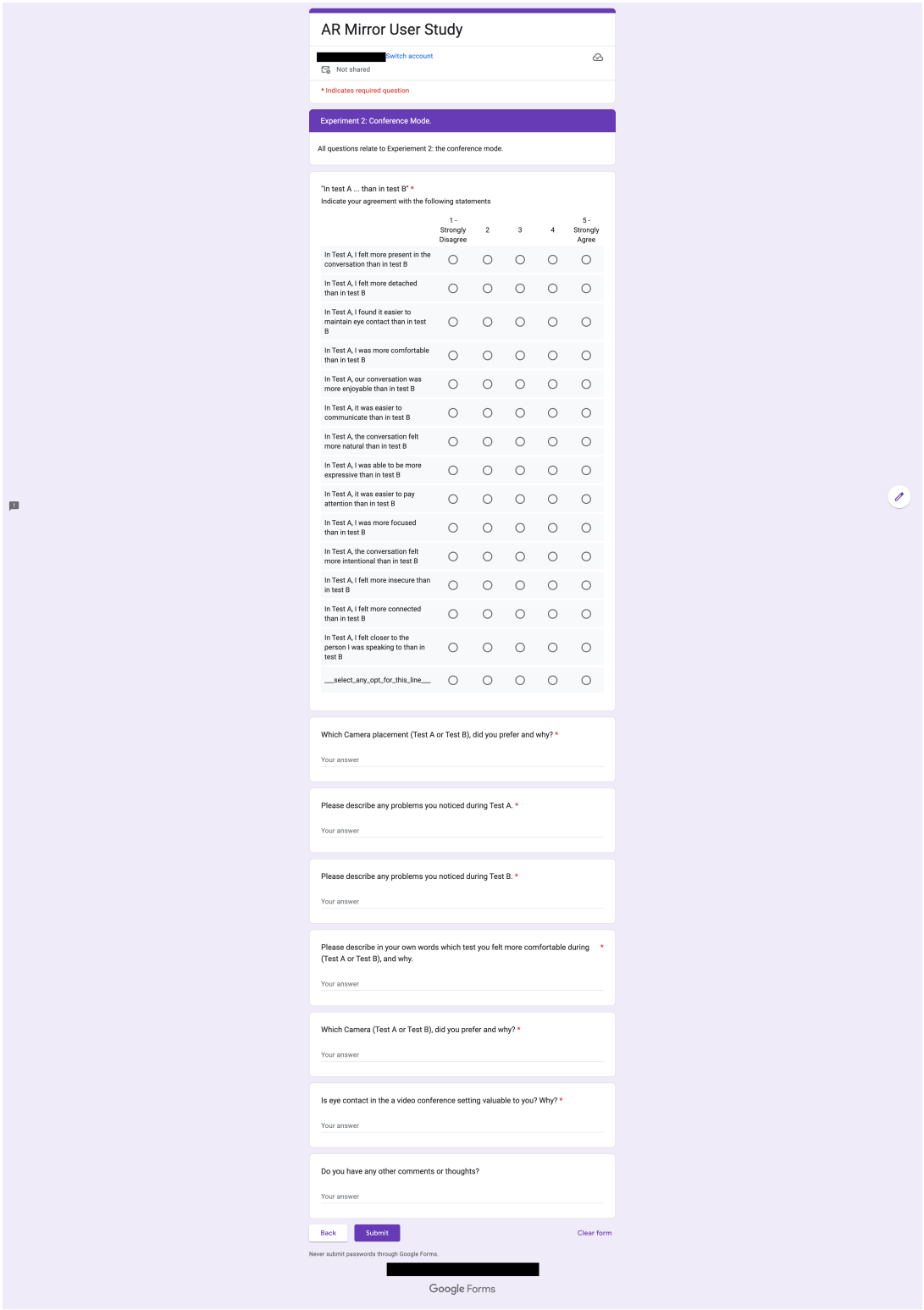

Fig. 7. User study screenshot (Page 3).