Co-Here: an expressive videoconferencing module for implicit affective interaction

David Marino* McGill University
Jiamin Dai† McGill University
Pascal Fortin‡ McGill University
Max Henry § McGill University
Jeremy Cooperstock¶ McGill University

Abstract

Participants of one-to-many videoconferencing calls often experience a significant loss of affective feedback due to the limitations of the medium. To address this problem, we present Co-Here: a videoconferencing module that renders participant-driven animations that convey group affect without relying on categorical emotion detection or signaling. Co-Here consolidates and visualizes facial expressions and head movements of fellow videoconferencers, providing a sufficient medium for others to meaningfully construct emotional impressions. Our qualitative user study showed that the system helped users feel a sense of alignment with the emotions of others using the system. Co-Here had the most utility to presentation viewers, who reported that the system offered a low-attention alternative to gauging the sentiment of the crowd. Co-Here further enabled a supportive environment that was described as between having cameras on and off. This encouraged users to emote more, and relieved social pressure commonly experienced in group calls.

Index Terms: Human-centered computing—Collaborative and social computing—Collaborative and social computing systems and tools; Human-centered computing—Visualization; Human-centered computing—Human computer interaction (HCI);

1 Introduction

A conversation is more than an exchange of words: it is accompanied by a wealth of paralinguistic, nonverbal, and contextual cues that give a rich interpretive medium for which layers of meaning can be derived [16]. Essential nonverbal cues, such as head motion and gesture, are often lost or degraded while using commercial videoconferencing platforms. This becomes particularly problematic when participating in a one-to-many videocall, such as a presentation or livestream, where there could be a total loss of audience feedback due to disabled cameras, a presenter taking up the majority of the viewport, or extremely low resolution of the viewers. In such cases, there can be a large loss of affective awareness as nonverbal behaviour such as facial expression are key to understanding the emotional state of interlocutors.

In the case of a one-to-many presentation, such as a conference talk, comedy show, or musical performance, it can be desirable for both viewers and presenters to have an understanding of the emotions of the people in the room. For example, a comedian may want affective feedback from the crowd to know if a joke they told landed, or a teacher may want to know if their students are bored or engaged. The audience, likewise, may want affective awareness of the crowd: in a concert, paralinguistic audience behaviour such as body movement, cheering, and entrainment with the performers, provides meaningful cues to enable an emotionally engaging experience [49]. A lack of such feedback can be a contributor to videoconference fatigue, as restricting the feed to the face alone places a cognitive burden on participants since they must use verbal information and limited visual cues to overcome a great loss of visual and nonverbal information [35].

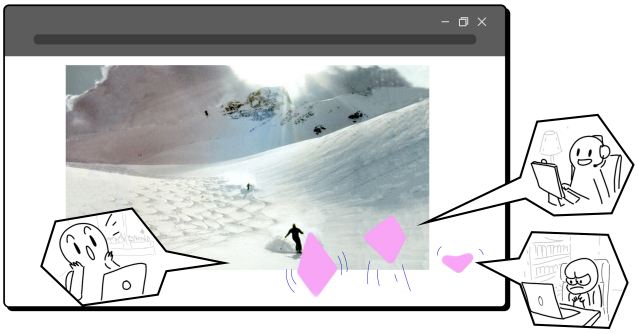

Figure 1: An illustration of Co-Here: the system displays the expressions of teleconferencing participants as a particle visualization to implicitly convey group affect during one-to-many calls.

We contribute Co-Here, a videoconferencing module that generates an abstract animation which is parameterized by audience facial expressions for 1:N video calls (conceptually illustrated in fig. 1). The facial expressions of audience members are mapped to individual particles ("particle-avatars") on the screen. A design goal of to Co-Here was to reintroduce aspects of paralinguistic communication experienced during live presentations to video conferencing. Many paralinguistic cues convey affective information, and are communicated passively without needing to use explicit symbols or inputs from interlocutors. We therefore aim to provide a medium for people to express themselves passively, much like how one would in a live presentation, without needing to rely on explicit signaling and classification of emotion. Additionally, we only utilize the user's webcam as input so that the system's sensing is non-invasive and accessible on commodity laptops. Through a qualitative in-situ study, we demonstrate that Co-Here provides an emotionally encouraging environment, inviting users to express themselves more freely and to a greater degree than they would in traditional large video calls. Co-Here can enable implicit emotional awareness during presentations without having to expose raw camera feeds, and without relying on categorical emotion signaling. We discovered that Co-Here can create emotionally cohesive presentations, inviting participants to "join in" on aligned emotional experiences.

2 Background

Co-Here approaches affective communication from a constructivist viewpoint. It differs from prior work by using this perspective to provide an ambiguous signal from which users can construct their own meaning. By doing so, it is a system that both *senses* and *conveys*

*e-mail: dmarino@cim.mcgill.ca
†e-mail: jiamin.dai@mail.mcgill.ca
‡e-mail: pe.fortin@mail.mcgill.ca
§e-mail: max.henry@mail.mcgill.ca
¶e-mail: jer@cim.mcgill.ca

emotion in an implicit manner for 1:N videoconferencing. A final distinguishing feature of Co-Here is that it uses non-invasive commodity hardware and displays affective feedback without sacrificing major screen real estate.

2.1 Affective Audience Sensing and Teleconferencing

A multitude of approaches have been used to sense audience affect. Audience emotion can be sensed by either *explicitly* asking audience members to select or signal their emotions or *implicitly* inferring through statistical models of behaviour and biosignals.

On the explicit side, Live Interest Meter was an app to convey audience engagement by explicitly polling audience members using their smartphones and presenting visualizations of the results [44]. Furthermore, the primary method of signaling emotion during presentations requires explicit symbols: emojis, reacts, and side-channel text are all examples of such methods.

On the implicit side, Biosignal sensing has a long history of use for understanding audience affect. Galvanic Skin Response (GSR) data were used to study the affective response of an audience to performing arts shows and live presentations [33] [52], as well as student engagement in distributed learning environments [51]. EngageMeter was a system that used electroencephalography (EEG) to sense audience engagement in conference settings and displayed engagement levels using a graph or scalar gauge visualization [23]. On the level of body analysis, posutural synchrony was utilized to infer affective feedback from audience members seated in ambient sensing chairs [53].

The human face offers a rich medium for affective feedback that has been often applied to teleconferencing. Simply compositing a feed of a remote conversation partner's face in the center of a user's gaze point, instead of the side of the screen, is sufficient to enhance feelings of emotional interdependence [28]. De Silva et al. [13] used a facial emotion classifier to animate exaggerated 3D avatars in a shared virtual space. Affective Spotlight is a Microsoft Teams extension that operates as a realtime video feed switcher by "spotlighting" user feeds that are algorithmically determined to be emotionally relevant while videoconferencing [38]. Using a similar video switching paradigm, motion detection and speech were used to cut between video feeds of a colocated meeting with the aim of enhancing engagement [43]. Multimodal face and speech data have been used in videoconferencing to classify user affect and display relevant emotion words over their video feeds [15].

During a traditional 1:N videocall, it can be attentionally demanding to to signal your emotions—one could manually select an emoji, or switch focus from the screen to the text chat. It is also has the potential to be attentionally demanding to other viewers as well—a deluge of side expressions in the chat window, or another user interrupting the speaker with emotive vocalizations can break the flow of conversation. It is thus desirable to have implicit input, similar to the prior implicit sensing devices reported in this section. This closely resembles what happens in real life, where there is no perceptual distinction between "input" and "output" of your emotions: you simply just smile to indicate some mental state without worrying about inputting the emotion to the system of your conversation. Prior work that utilized specialized devices such as EEG or GSR are invasive for everyday use, or require specialized hardware not commonly available to users. In constrast, Co-Here uses everyday sensing devices (i.e., the user webcam) to implicitly translate affective signals. This approach to emotion sensing is not unique in of itself, but the combination of how Co-Here both senses and represents emotion, and how that is applied to teleconferencing, is what distinguishes it most from prior work.

2.2 Representation of Affect

All affective teleconferencing systems presuppose a theory of emotion, but emotions are ontologically difficult to define. Many systems that use emotion classifiers operate on the assumption of basic emotions—defined as a set of discrete, psychologically primitive affective states. Some basic emotion theories posit that there are universal atomic emotions shared between cultures [50]. Many emotion classifiers aim to map a biophysical signal to a primitive emotion, which is then typically presented to the viewer [46]. A popular basic emotion theory posits that there are seven universal facial expressions, known today as "Ekman faces" [17]. Despite its widespread theoretical adoption in AI and HCI research, the notion of universal facial expressions has little empirical support as facial configurations have been shown to map to multiple emotions, and vary across cultures [5]. Additionally, an agent's emotion cannot be be wholly determined by a single signal, such as a smile, or vocal quality; the signal is always situated in a complex context which affects its interpretation [9]. Such an interpretation of emotion is consistent with a constructed theory of emotion. In this theory, emotions are not construed to be atomic universals but instead concepts that arise by utilizing the brain's inherent pattern-generating capabilities, integrating past experience and realtime ambiguous stimulation [3, 4]. We adopt this theoretical position when designing our system. Our system does no explicit emotion detection. Instead of classifying discrete emotional states, or showing users a representation of atomic emotions, we aim to show the user a sufficiently ambiguous signal for which they can ascribe emotional meaning to themselves. By doing so, we shift the burden of emotion classification from the computational system to the user, and utilize the brain's ability to form patterns and concepts from ambiguous data [48].

We are of the position that it is important to represent emotion implicitly during a video call, because that is how emotion is represented in real life. Explicitly signaling emotion can break the flow of conversation, and has the potential to divert attention from the topic at hand. It also runs the risk of misrepresenting what users feel because it suffers from low resolution. For example, consider if a user smiled during a conversation, and the system subsequently signaled "HAPPY". Perhaps they were only slightly happy, perhaps they smiled to make others feel more at ease, perhaps they smiled because they were uncomfortable. Regardless of the original intent, viewers would have a unified yet inaccurate picture of the source emotion. There are also times when basic emotions are of an inadequate granularity. For example, one study prototyped a 1:1 videocalling app for individuals with Autism Spectrum Disorder where facial expressions were translated to explicit symbols such as emojis or emotion words based off six basic emotions [6]. Participants found basic emotions unhelpful because they struggled more with understanding complex emotions such as frustration, sarcasm, or confusion [6].

Co-Here thus strives to convey emotion to users without using explicit affective symbols. In terms of a design philosophy, Co-Here is most aligned with that of affective interaction, where emotion is considered inherent to interaction—it is "dynamic, culturally mediated, and socially constructed and experienceed" [7].

2.3 Backchannel and Grounding in Conversation

A grounded conversation is one where interlocutors continuously coordinate to establish shared common knowledge and beliefs [11]. Grounding in conversation is important not just for successful dialogue, but also as a necessary sociotechnical condition for effective remote work [11, 40]. A closely related concept, that of interactive alignment, extends the notion of grounding—it claims that when successfully communicating, interlocutors align representations among all levels of language: phonetically, syntactically, semantically, and situationally [19]. This is evidenced both behaviorally by shared linguistic constructions between conversants, and also neurobiologically in studies that show a coupling between perception and action between conversants completing joint conversational tasks [20, 36, 54]. In the field of human-robot interaction,

grounding has been extended beyond conversation to also encompass affect. This has been called *affective grounding*, where affective ground is established when interactors coordinate on how behavior is to be emotionally understood [27]. Just as grounding in communication conceptualizes conversation as collaborative, affective grounding conceptualizes emotion as collaborative.

Videoconferencing poses challenges to grounding: interlocutors do not share the same physical environment, and sometimes they are not visible [11]. In 1:N communication in particular, the challenge to grounding is even greater as obtaining a shared situational sense between all conversants may be near impossible. A crucial mechanism of coordination that enables grounding is backchannel communication—the presence of verbal and nonverbal cues like "mhmm" and head nods during conversation [27, 42]. However, in a video call, backchannel communication is often heavily suppressed. A high level design goal of our system is to facilitate the establishment of affective ground between interlocutors. We primarily tackle this problem by visualizing the non-verbal backchannel of head motion and facial configuration of participants.

3 System Design

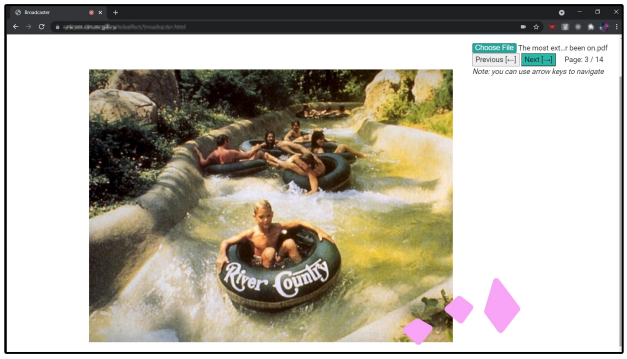

Figure 2: A live screenshot of the GUI with 3 participants. The pictured stream is from a slideshow with a high valence, high arousal affective target.

Co-Here is a browser-based videoconferencing system that communicates the expressions of participants to one another with the goal of enhancing affective awareness. A screenshot of the GUI is presented in Fig. 2. Our code is publicly available on GitHub, linked in the footnote.[1] The system is currently designed around the task of giving 1:N presentations, in the form of slideshows. This first application was chosen since it is representative of numerous 1:N remote activities, e.g., teaching and academic presentations. Furthermore, since a slideshow typically takes up the majority of the viewport, it imposes itself as a scenario where the loss of non-verbal cues from the audience can be particularly severe. Facial landmarks were extracted from user video feeds, and utilized to animate the particle-avatars.

A single particle avatar is animated as follows: absolute positioning of the head affects the avatar's origin point; head roll, pitch, and yaw move the avatar in corresponding directions; eyebrow motion adjusts the top size of the avatar; mouth y-distance affects the bottom size of the avatar; mouth x-distance adds sinusoidal motion and expands the sides of the avatar. These mappings are summarized in Table 1. A sketch of the algorithm showing the relationship between facial landmarks and particle avatar parameters is outlined in Fig. 3. While there were a multitude of animation possibilities, the current methods were chosen to convey a natural mapping between

[1] `https://github.com/Shared-Reality-Lab/co-here` (US 18/463,799)

Facial feature	Animation
head position	particle location
head roll, pitch, yaw	particle displacement
eyebrow motion	height of top of avatar
mouth vertical position	height of bottom of avatar
mouth width	sinusoidal motion

Table 1: Mapping of facial features to avatar animations.

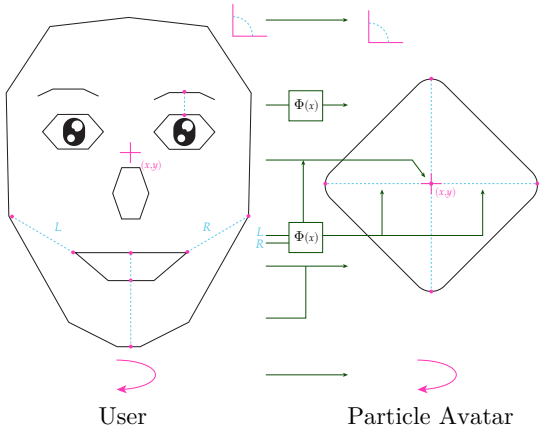

Figure 3: Animation parameter mapping between a user and their particle avatar. Lines with $\phi(x)$ utilize nonlinear mappings, where plain lines utilize linear mappings.

live head motion and particle animation. An author who is also an experienced motion graphics designer manually tuned particle avatar animation parameters through trial and error to create animation they found to be compelling. Landmark data was captured at 14 fps and filtered using an autoregressive moving average function to smooth the signal.

New users must first calibrate their facial parameters prior to using the system. This is done automatically by linearly interpolating participant faces to min/max ranges that are ideal for compelling avatar animations; however users are also given slider control to manually adjust animation settings. This was because the "basic face" used to automatically normalize participant face parameters was that of one of the authors. Since no two people share the same resting face [34], the additional manual controls were provided to ensure that the particle avatar was sufficiently responsive to each user.

When a presentation starts, a number of particle-avatars populate the bottom right of the screen. Every participant of the video call, including the presenter, has a corresponding particle avatar shown on screen. No user video data is shown to other participants, stored, or transmitted. Co-Here is designed to augment a traditional videoconferencing system by overlaying particles to a video feed, i.e., of a presentation or screen share. All avatars are of uniform color and size for a sense of anonymity. A high level overview of the system is shown in Fig. 4.

4 Methods

A user study was conducted to investigate the experience and feasibility of using such a system to convey affective feedback while videoconferencing. The system was evaluated qualitatively, focusing on a live presentation task with groups of 3-5 concurrent users. The study was approved by our institutional research ethics board.

We investigate the following research questions (RQs):

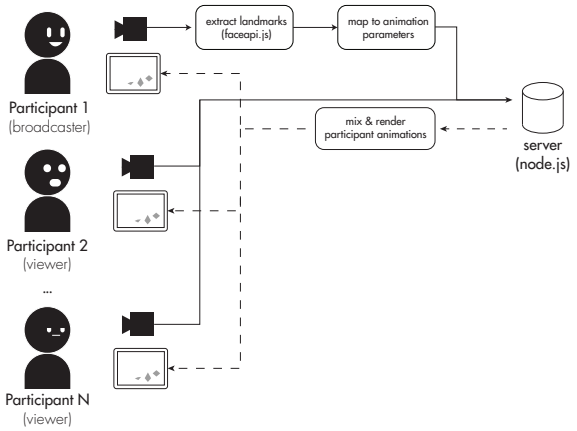

Figure 4: High level system architecture. Participant face landmark data is extracted from their webcam feed. Analysis is conducted to map landmark values to animation parameters for their particle avatar. A server broadcasts all participants' animation parameters. Values are then mixed together and animated onto the screen for all users.

PID	Videoconferencing Experience					Profession
	Friends/Family	Work	School	Performance	Conferences	
P1	1	1	0	0	0	comedian, copywriter
P2	1	1	0	0	0	ux researcher
P3	1	1	0	0	0	editor
P4	1	0	0	0	0	-
P5	1	1	0	0	0	programmer
P6	1	1	0	0	0	speech language pathologist
P7	1	1	1	0	1	grad student
P8	1	1	0	0	1	grad student
P9	1	1	0	0	0	-
P10	1	0	0	1	0	-
P11	1	0	1	0	0	student
P12	1	1	0	0	0	writer
P13	1	1	1	0	0	grad student
P14	1	1	1	0	0	undergrad student
P15	1	1	0	0	0	architect
P16	1	1	1	0	0	grad student
P17	1	1	1	0	0	grad student
P18	1	1	1	0	1	undergrad student
P19	1	1	1	0	0	grad student
P20	1	1	1	0	0	highschool teacher
total	20	17	9	1	3	

Figure 5: Participant background information based on their prior videoconferencing experience and current profession. Participants were experienced videoconferencers: all had prior videoconferencing experience with friends and family, with most also using videoconferencing at work or school.

- **RQ1**: How do users understand the meaning of Co-Here's animations?
- **RQ2**: How does Co-Here phenomenologically differ from a traditional video call?
- **RQ3**: What types of interactions does Co-Here best facilitate, and what are the shortcomings of such an interaction technique?

4.1 Participant Background

We recruited 20 participants through a combination of social media ads and snowball sampling. Each study session started with an icebreaker activity where participants were asked to introduce themselves and say what they used videoconferencing for. Participants were also given the option to reveal their professions. Not all disclosed this information. Participants' videoconferencing background is listed in Fig 5. All participants reported that they had prior experience with videoconferencing. The most common use cases were for talking to friends and family, followed by work, school, academic conferences and performance. Here "performance" means entertainment-based 1:N presentations such as improv or comedy shows.

4.2 Study Procedure

Participants were asked to use the system to give presentations to a live audience. We hosted 5 sessions, each containing 3-5 participants. A session consisted of the following phases:

1. Icebreaker activity: participants introduced themselves to one another.
2. Presentations: participants took turns giving 2-5 minute presentations.
3. Semi-structured focus group: guided questions at the conclusion of the presentations to discuss the experience of using the system.

We asked participants the questions: "How was using this system in comparison to your everyday videoconferencing experience?", "Were there moments that stuck out to you in terms of how people were feeling and reacting to the presentations?", "How did the system affect your attention during presentation?", and "What do you consider the ethics of using such a system to be in everyday use?". These questions were designed as starting points to create a greater conversation for which more precise and circumstantial follow-up questions could be asked.

To fully understand the expressive limitations of the system, presentations were designed to cover a large emotional range by targeting quadrants of Russell's 2D circumplex model of affect [45]. Russell models affect in terms of two dimensions: *arousal* which describes levels of alertness, and *valence* which describes a range of pleasure and displeasure. Taking combinations of each quadrant, presentations were crafted with the following affective targets: {HIGH AROUSAL, HIGH VALENCE}, {HIGH AROUSAL, LOW VALENCE}, {LOW AROUSAL, HIGH VALENCE}, {LOW AROUSAL, LOW VALENCE}, {NEUTRAL AROUSAL, NEUTRAL VALENCE}. To create a presentation, a single researcher selected validated images from the International Affective Picture System (IAPS) that were consistent with the affective target [31]. The goal of selecting validated images was to ensure that the affective target was reached at least once in a presentation. Images were sequenced to lay the foundations of a 2-5 minute slideshow story. Images from the database that were subjectively deemed to be inappropriate for day-to-day teleconferencing (e.g., explicit sexual content or gore) were excluded. Filler images to "complete" the story were sourced from Google Images. Natural, compelling stories and presentations have dynamic and changing emotions. Thus, a single slideshow was not assumed to be emblematic of a single emotion. The slide shows were designed to be ambiguous to leave room for improvisational user interpretation. As such, they incorporated little to no text and the connections between successive images were not explicitly stated. Participants were given the option of improvising a story to go along with the slide show, creating their own notes beforehand, or, if they were uncomfortable, asking the researcher for a script. For the participants who asked for a script, a single researcher wrote one stream of consciousness to the selected images. Participants were also given the option to modify the slide show in any way to serve their version of the story so long as it was consistent with the affective target (i.e., the slideshow contained validated images that were consistent with the affective target). One participant opted to include their own photos.

In a single session, slideshows were assigned to participants in a way such that there were no duplicates of affective targets. Over the course of the entire study, 45% of slideshows had positive valence targets, and 50% of them had positive arousal targets.[2]

4.3 Qualitative Investigative Technique

We conducted an inductive content analysis of focus group transcripts for this exploratory investigation [10, 14]. We followed the technique outlined by Hsieh and Shannon which provides a "bottom-up" approach to group raw data to overarching themes [26]. Our method was as follows: first a single researcher researcher coded interview segments according to its literal meaning. The codes were meant to be descriptive with little interpretation, often including simple key words. Afterwards, two authors reviewed the transcripts and codes, and clustered quotes into "categories" based on similarity of meaning. The categories are largely summative with little interpretation, focusing on the "who", "what", "where", "when" of groups of codes. Finally, the same two researchers further clustered the categories into themes. Our themes describe overarching inductive meaning, whereas the categories mainly describe deductive meaning. In this manner, our themes present the "why" and "how" of our data. They then discussed and merged the categories and themes until reaching an agreement on their interpretation in relation to our RQs. Our qualitative data is publicly available in an anonymized form in the footnote [3].

5 Findings

Our content analysis revealed seven categories across two themes (Fig. 6). We include a breakdown between participants and the categories they were associated it in Fig. 7. Below we elaborate on each theme and category, and conclude with a review linking the themes to our research questions.

5.1 Theme 1: A social space between mediums

Participants felt that Co-Here presents a unique social space, somewhere between a phone call and a video call, and between having cameras on and off (1B-Interstitial Living Experiences). The anonymous nature of the space helped to relieve pressure and encouraged participants to emote more freely (1C-Promotion of Emotional Expression). For viewers, it provided a lower attention alternative to assessing the feelings of the crowd, but this sentiment was not shared by presenters (1D-Attention). Co-Here provides an empathetic experience, which encouraged participants to align in emotions with one another (1A-Alignment). Co-Here presented a unique kind of telepresence. It created a distinct social space between various modes of online communication, inviting interactions from participants distinct from traditional videocalling. Below we show each category that emerged from content analysis, and conclude on how the categories relate to themes.

5.1.1 Category 1A: Alignment

The alignment category describes phenomena where the emotions or behaviour of participants synchronized. Particularly, many participants had the urge to match their facial expressions to the perceived emotions of the audience, as to act in an emotionally and socially unified way. P9 felt a sense of emotional unison with others, saying "I felt encouraged to want to join in on that same emotion when seeing those visual cues".

P7 and P3 both wanted to react in the same way that others in the audience were reacting, especially when it came to positive affect.

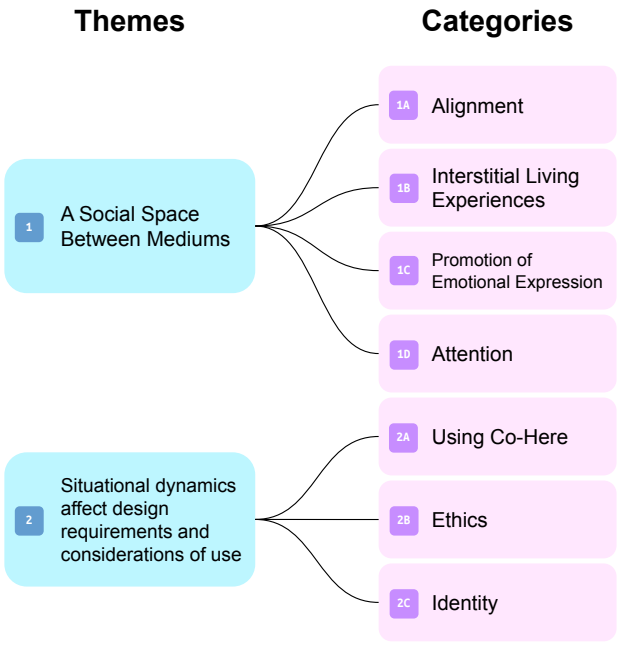

Figure 6: An overview of the themes and categories that emerged from Content Analysis. The themes are overarching interpretive topics that are comprised of multiple categories.

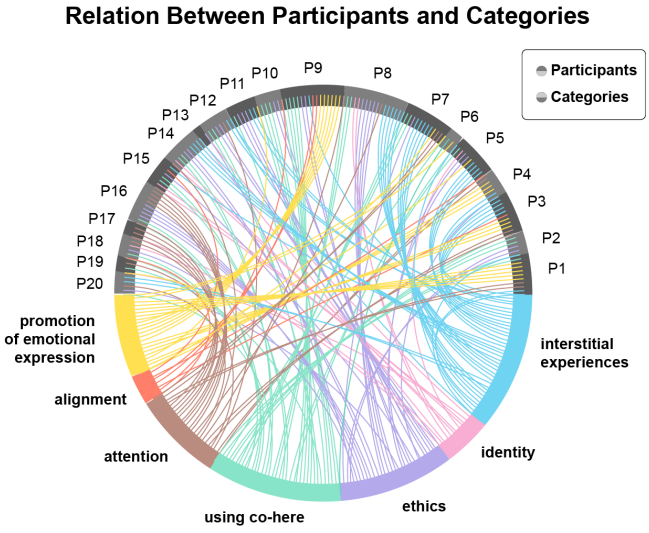

Figure 7: A breakdown of the categories associated with participants. The top semicircle represents participants, and the bottom semicircle represents categories. Each chord is a quote from a participant that has been categorized. For example, P1 has 11 quotes that were divided into 4 different categories. The width of the categories show the proportion they occurred in our corpus. The width of the participants show how many quotes they contributed to our analysis.

> "I kind of felt influenced to match or replicate what I thought other people were doing, like a big smile or something like that". –P7

> "One thing I found very interesting is that I noticed other people reacting a certain way and it made me feel like I

[2]Despite having 20 participants, these numbers were not perfectly 50/50 due to participants canceling/dropping out and new slideshows having to be scheduled ad hoc.

[3]https://airtable.com/appTGhYR1opVIzBgw/shrQXOEJpCHLiWQmc

wanted to react that way, I was like everyone's smiling I should also smile" –P3

There was also meta-awareness of how other people's emotions affected their own perceptions of the presentation material. For example, a negative valence, high arousal presentation had a jumpscare, but it caused a chain reaction of laughs. P15 remarked on this, saying:

> "There was like this creepy creature and someone laughed and it made me laugh". –P15

P5 became aware of how their own emotions may affect other people's, and recognized it as a feature that made them feel a part of a community.

> "[When] I smile and [it makes] somebody else smile it does make it feel like you're in the community space a little bit or like I guess approximating towards that". –P5

There was also individual concern if one's own emotions matched the crowd. P19 said that while watching presentations "I was seeing, is everybody else laughing as well, or just am I?". P10 shared a similar sentiment, wanting to socially contextualize their emotions, saying: "I was like oh this is humorous to me, and [then] I was like 'oh I wonder how other people are reacting to it'."

Co-Here thus enabled participants to affect each other's emotions and behaviour in a passive manner.

5.1.2 Category 1B: Interstitial Living Experiences

Most participants reported feelings of "inbetweenness" when using Co-Here. These feelings were described from multiple angles: one aspect was in the participant's relationship to their cameras and privacy. A few participants described Co-Here as "in the middle" of having one's cameras off and on.

> "[I] see it being a nice medium between total video off and on." —P8

> "[It is] a nice in-between... [where] you want people to know that you're present and listening but you're not super comfortable turning on your camera on" —P14

Another aspect of inbetweenness was related to paralinguistic aspects of communication, such as backchanneling (concurrent feedback from interlocutors), gesture, and side communication. During videocalls, conversations tend to be more turn-based, and aspects of communication that occur *between turns* tend to be reduced, thereby discarding many paralinguistic aspects of conversation. Users expressed that Co-Here brought back some paralinguistic aspects of communication, particularly a greater ability to communicate emotional information between conversation turns.

> "I think it's almost unanimously agreed that speaking in a [video] conference always feels like you're interrupting somebody... there's no room for sharing a moment or sharing an emotion with somebody who's not necessarily speaking. So being able to have this little 'oh we're communicating' like little blobs and it's like smiling... make it feel like you're in the community space a little bit or I guess approximating towards that" —P5

Finally, another aspect of inbetweenness was related to embodied vs virtual experiences while teleconferencing. For example, P4 said that the visualizations felt distinct from the disembodied experience of signaling emoticons from one another, and the way they responded to the audience made the visuals "feel like it's alive". Similarly, P11 expressed:

> "It replicates a little bit more giving a presentation in real life just because when it's over a video call, it's so zoomed in on people's faces, it's not like that, they get blended into a crowd." —P11

5.1.3 Category 1C: Promotion of Emotional Expression

Many participants felt that the feeling of in-betweeness took the pressure off of communication and by doing so encouraged them to emote more.

> "It took a lot off a lot of the pressure on my reactions so I could express freely. I feel like there's like this unspoken code where you shouldn't laugh too much or you shouldn't look too pissed off... but if you're anonymous and you can't hear them you actually express as much as you want and I think there's a freedom to it." —P9

These sentiments were also accompanied by reports of a reduction in anxiety. P11 said that the lack of a live video stream takes off pressure for both viewers and broadcasters, and because of that, you get more genuine reactions.

Asides from a reduction in anxiety, the particle avatar control system itself encouraged users to express themselves in unique ways. Some users felt that just being able to control an avatar encouraged them to express themselves more.

> "I like having something that I can control on the screen, so I like [that] as soon as I know that I can display emotions and react [it] makes me want to react more and engage more" —P5

The particle-avatars do not have the same degrees of freedom as a real face. As such it is a low resolution approximation of a facial expression. But because of this, users were encouraged to express themselves in ways that were best suited to the particle avatar's animation parameters.

> "I wanted to it to be more emotive you know, sometimes when I was really reacting I get really close to the camera then my [avatar] would get really big just because I was like 'I need them to know like that I'm really happy about hearing this'... I was trying to kind of push it to the limits to... show the presenter [how] I felt about it" —P4

P1, who is a comedian, brought up that emoting is often a form of "support", and that at a comedy show emoting in the audience is a form of supporting the comedian. P9 felt this sentiment while using the system, saying that

> "as a presenter it was also almost a like a subtle encouragement to see people reacting on the side... [and] it takes off the pressure when everyone's anonymous, not seeing everyone's faces made me feel less like the eyes were on me". –P1

5.1.4 Category 1D: Attention

This category touches on aspects of the system that both negatively and positively affected the subjective experience of participant's attention.

Co-Here showed potential in being a low attention alternative to traditional video tiles on the basis that it aided in reducing distractions and mental workload. One source of distraction from traditional video calls can emerge from unshared aspects of interlocutor environments. For example, P15 said they always felt distracted by the camera feeds of fellow videoconferencers due to wanting to microanalyze everyone's backgrounds, with P3 further saying:

> "I do think that this is less distracting than having cameras on because... I feel like when I go to big meetings I go through people's photos see what [they're] doing, like what room are they in... it's more information that way so this is to me less distracting". –P3

Another source of distraction comes from anxiety around other people scrutinizing each other's faces, including their own.

> "I prefer [Co-Here] over having actual video feeds of myself and of people. I find video of just people's faces really distracting and when I'm a viewer I end up getting really concerned with how I look". –P16

That said, there were several ways Co-Here could potentially be detrimental to user attention. One is due to a novelty effect that some participants found distracting.

> "It's still a new experience... my immediate association with it is that talking mirror from Shrek because it looks like that and it moves like it and so it feels cartoony... it would take practice with it to understand it, and not to be so distracting". –P9

P3 further said that they felt it was distracting at the start because "I'm trying to see everything I can do with it", though later in the interview reported that in comparison to traditional videoconferencing, they found it less distracting.

Another way that Co-Here could be detrimental to user attention comes down to the role users had during presentations. Viewers tended to report feeling less distracted than presenters. Many presenters reported not to look at the visuals, opting to focus more on their notes. P18 said that while broadcasting "I gave less attention to [Co-Here] because my focus was more on speaking stuff and getting everything on track". P16 said that while they were a presenter, they only noticed the visuals when they were trying to make a joke, or say something with emotive impact.

There is also an effect of familiarity with the presentation material itself, which affects attention: P1, a comedian, said that as a presenter he didn't look down at the visualization much because he wasn't familiar with the material and didn't have expected reactions from viewers. He emphasized that if it he had material that he knew intimately it would greatly impact his relationship to Co-Here because he would be looking for specific reactions.

Of note is a distinction between presenters who decided to improvise their lines, and those who pre-wrote scripts. Improvisers were split on how "tuned in" they were to the visualizations. P6 found much utility in the visuals, stating:

> "As a presenter I was pretty tuned in [to the visualization] and I was like oh no I hope they're not like dead pan or like tough crowd or whatever". –P6

Contrarily, P10, who also improvised her lines, did not attend to the visuals citing issues of familiarity:

> "When I was giving the presentation, part me was kind of like oh it would be interesting to see if it helps gauge people's reactions, but I found that because I was so focused on being 'okay wait what slide is going to come up next' I didn't find myself looking down and kind of knowing how people react". –P10

5.2 Theme 2: Situational dynamics affect design requirements and considerations of use

Co-Here was designed for 1:N video calls, but relevant nonverbal cues vary widely between different videoconferencing contexts. Co-Here's design, including its anonymous nature, created a specific social environment. This theme discusses how that environment relates to real world use, the effect of the design decisions behind Co-Here, and how the system may change in the future. We discovered Co-Here's appropriateness to current and future use cases (2A-Using Co-Here), aspects around the nature of anonymity while using Co-Here vs traditional videoconferencing (2B-Identity), as well as ethical dimensions to the system in use (2C-Ethics).

5.2.1 Category 2A: Using Co-Here

This category discusses how the context of use affects design requirements. It particularly covers topics of what type of videocalls Co-Here is most appropriate for, and how the design of the system itself could change depending on different use cases.

Participants unanimously agreed that Co-Here is appropriate for 1:N presentations. Most critically, it was discussed that Co-Here should be used for emotion-centered presentations.

> "If you're giving a serious presentation where emotion is not really part of the picture then I feel like this won't be as useful vs if you're doing a twitch stream or something and emotions are a big part of that ". –P2

The notion of "emotional support" or "encouragement" was at the crux of determining Co-Here's situational appropriateness. In cases where emotional support isn't needed, such as in some specific work meetings, Co-Here was seen as not being as useful. Commenting on this aspect, P1 simply said "there's some cases where everybody... is muted for a reason". Conversely, P5 felt that the device was helpful in keynote talks "to encourage the speaker" since these are professional settings where emotional feedback could be encouraged.

A major concern of using Co-Here in professional contexts was the amount of intentional control users had over their particle-avatars. In professional meetings, sometimes tight control is desired over the emotions you express, and a highly interpretive emotional visualization such as Co-Here may have unintended consequences. P9 told an anecdote of giving a presentation in a work meeting, where a client yawned and it threw them off.

> "[The yawn] immediately sunk my confidence, which wasn't very much to begin with... but if we were to have a platform like this I think I would have preferred not to have seen him yawn. I would have preferred him to have his own little avatar and reacting however he wants and he can convey whatever emotions or questions that he wants behind a mask". –P9

Non-professional contexts, such as watch parties, were more commonly accepted environments to use Co-Here.

> "For something like a watch party where it's low stakes... it's a nice way to just get a sense of 'yeah I want to feel what the vibe is'.". –P2

Education (for students) is a pseudo-professional, "in-between" use case that generated a lot of debate among users. The general feeling of "taking the pressure off" was seen as being valuable in a classroom environment. P20, who is a high school teacher, saw that the anonymity afforded by Co-Here could be supportive to his students needs. While teaching during the pandemic, he remarked:

> "Students just like to be hidden... [maybe] they don't want to show their messy room... [maybe] they just like that anonymity... and I think there's just a lot that has to do with self esteem". –P20

P8 added that there are also circumstances where students wouldn't want to have their cameras on because of concerns about publicly sharing aspects of their living environment or family, so a visualization such as this is an ideal medium. P14, a business and computer science undergraduate student, felt "it's another way for students... to be more interactive in even a lecture based format" but stressed that there were different degrees of interactivity required between their computer science and business classes, and that emotive visualizations are best suited for more interactive classes. However, P6, a masters student in computer science, said that there are times when she simply prefers to be fully hidden in class: "sometimes in lecture, I prefer [to have] camera off because I can listen to it like a podcast, and I don't have to be hyper attentive, and it helps with Zoom fatigue".

Many different presentation types and use cases have been discussed, but each unique activity may have different design requirements. P12 highlighted the fact that relevant nonverbal feedback differs between presentation types, and thus may require different animation mappings:

> "[The] kind of data I'd be looking for might be different to if I was in a work meeting, and it was a collaborative project and I was looking for people shaking their head or nodding" –P9

There are also some situations where the use of the visual modality for affective feedback is not as useful. As previously mentioned in Category 1D: Attention, the high visual mental workload when presenting meant that some broadcasters didn't have the capacity to attend to the visualizations, and thus the audio modality was suggested to represent affective audience feedback instead since that is how it is typically experienced in colocated performances. P16 brought up an example of watch parties, saying that they would personally prefer to receive affective feedback auditorily as to not crowd the movie screen. The system as it is presently designed is thus concluded to be best suited for informal emotion-centered 1:N casual presentations such as twitch streams, or informal conferences.

5.2.2 Category 2B: Identity

There was much debate among participants as to the value of having anonymous feedback. While many participants agreed that anonymous feedback helped alleviate pressure while videoconferencing, there were some who were nonetheless very concerned with identifying who they were. P4 and P8 both mentioned that they would move their faces in exaggerated manners to try and identify themselves in the crowd. P19 found themselves very preoccupied trying to link particular avatars to people. P14 had a differing opinion, saying:

> " I feel like there's more of a novel underlying idea behind keeping it anonymous because... if you keep it anonymous then it's easier to find a genuine reaction and it's easier to be confronted with a genuine reaction of your audience." –P14

Indeed, participants who felt that the system encouraged greater emotional expression attributed it to the anonymity the system afforded, and consequently, a release of social pressure. P13 offered a design solution for those who were preoccupied with identifying themselves without sacrificing anonymity: privately highlight your personal avatar, while keeping the others anonymous. This will be considered for future iterations, as implementing this could conceivably aid in reducing the novelty effect by reducing the amount of time it takes to find yourself in the crowd, as well as assist users who have a propensity to monitor themselves in video calls.

5.2.3 Category 2C: Ethics

Participants were asked, very generally, of ethical concerns they had of the system's use. Many agreed that the device should follow standard "zoom consent rules" by explicitly asking permission. P11 said: "there's more importance of our consent because it's collecting biometrics data " (i.e., landmark extraction). P11 stressed that there was a fundamental distinction between a raw video feed that you would encounter on Zoom, and Co-Here, because Co-Here conducts analysis on the video feed. Some participants had used similar technology before (e.g., a Snapchat filter) and had relaxed feelings about using Co-Here so long as it was optional and for casual use. P11 said "I would be comfortable using it for entertainment purposes, so like [P12] mentioned like Snapchat, I consent to that and that's totally fine. But I would feel weird about it if it was like required by my school or required by my work". The importance of choice was further illustrated by P5: "I think it's all about choice. If there's an avatar meeting, it's not obligatory to have it, you should be able to choose that". P16 and P18 had concern about facial data being stored, because then it could potentially be used to conduct further analysis, to which they did not consent. Currently Co-Here does not write facial data to persistent storage; rather, the server keeps landmark data in RAM for only 70ms after which it may be garbage collected.

Other ethical concerns were centered around hypothetical oppressive design choices such as if Co-Here was modified to use for attention tracking by detecting if someone was looking away from the screen. Attention tracking is a timely and relevant concern of the system since it has been used in many commercial videoconferencing platforms such as Zoom, and WebEx, mainly using window focus as the primary mechanism to track user attention.

5.3 Connecting Themes and Categories to Research Questions

5.3.1 Research Question 1—How do users understand the meaning of Co-Here's animations?

We learned the following about how users understand the meaning of Co-Here animations (**RQ1**): Users understood that they could directly influence the particle-avatars, and would think about how to control the animations themselves to imply certain emotions to others. They likewise understood the animation as being "controlled" by others, and were trying to think of ways to play social metagames, such as expressing encouragement, agreement, or specific reactions to events. Users tried to align their affective expressions to what they perceived other audience members were feeling, which required some understanding of the meaning behind the animations . In this way, users behaved as if they were acting in a social space, where alignment of expressive behaviour is a common occurrence [19] Nonverbal feedback conveyed through Co-Here affected the participants' own emotions, and viewers felt encouraged to join in and also direct the emotions of others through their facial expressions.

5.3.2 Research Question 2—how does Co-Here differ phenomenologically from a traditional video call?

Co-Here occupied a space somewhere between having your camera on and off in a video call. There are many times when keeping the camera on during a video call for an extended period of time can result in fatigue—popularly "zoom fatigue". Co-Here offered users an opportunity to deliver anonymous continuous feedback, which is not possible in a traditional video call because cameras must be turned on or emojis / text must be explicitly signaled from your account. This is exemplified best by two participants who felt Co-Here enabled "side channel" communication where you can passively gesture to one another without interrupting the flow of conversation. Co-Here was experienced as being less stressful than a traditional video call while maintaining emotional engagement. Six participants felt that the system encouraged them to express themselves

more—both because of the nature of the visualization system and also due to its less stressful and anonymous environment. Of note are participants who reported they felt "supported" by seeing audience feedback, yielding a friendlier, less stressful videoconferencing environment.

P11 made the strong claim that presenting on Co-Here felt closer to real life than in a traditional video call, because the animation mixes a crowd of people together instead of presenting a tile of video closeups.

5.3.3 Research Question 3—What types of interactions does Co-Here best facilitate, and what are the shortcomings of such an interaction technique?

Co-Here was used to offer unobstructive side channel communication in an implicit manner (P5). While traditional videoconferencing platforms offer methods for side channel communication (e.g., direct messaging), all of these channels can interrupt the flow of communication and require explicit focus. On this note, Co-Here was considered to be "low attention" mainly by the viewers, but not the presenters, with the only exception being one presenter who opted to improvise her lines. The presenters largely ignored Co-Here visualizations with the exception of improvisers and moments where reactions were expected. Co-Here in its current form is thus best suited for viewing presentations. As seen in colocated presentations, affective feedback is also given in non-visual modalities such as sound when laughing. Future versions of Co-Here could use different modalities to deliver affective feedback to presenters.

We conclude that Co-Here best facilitates remote 1:N casual or improvisational presentations, and is most for viewers to get a "vibe check" of the audience. It currently is inadequate for presenters, and would require further design iterations to address this shortcoming.

6 Discussion

6.1 Emotional Alignment & Contagion

An unexpected finding while using Co-Here was the emotional alignment experienced by users of the system. Alignment of emotion is popularly associated with the psychological concept of emotional contagion where the emotions of one can "spread" to others. Emotional contagion is known to happen in various forms of computer mediated communication (CMC), such as in instant messaging—or, perhaps most infamously, through social media [22, 29]. Emotional contagion has also been observed in telepresence robots, dyadic video conferencing, and 1:N livestreams [8, 21, 30]. Was emotional contagion present while using Co-Here? This is a question best answered experimentally. It is an open question whether the emotions that "spread" to other participants were as consistent as they are in real life conversation, and an experiment is needed to demonstrate a causal relationship between individual initiators of a particular emotion that is then "caught" by recipients. If contagion did indeed occur while using Co-Here, it would be a unique channel for it to arise in videoconferencing. Raw video feeds or text chats are the typical vectors of transmission when analyzing emotional contagion in videoconferencing [21,37]. Here, participants reported the feeling of their emotions align with members of the audience based purely off a novel visualization.

Co-Here could have enabled contagion in a variety of ways. Emotional contagion has been most popularly hypothesized to happen in the following steps: mimicry, feedback, and finally "contagion" [25]. Mimicry as a basis for emotional contagion is perhaps the most popular theoretical framing of the phenomenon [18, 24]. Some theorists would call this a "primitive emotional cotagion" which is "relatively automatic, unintentional, uncontrollable, and largely inaccessible to conversant awareness" [24]. In the Mimicry step, Hatfield et al. observes that humans have a subconscious tendency to mimic facial expressions, vocalizations, and posture. Next, the feedback step posits that feedback from our own expressions (facial, vocal, postural, or otherwise) modulate our experience of emotions [25]. This concept has roots as far back as Charles Darwin, who hypothesized that our facial expressions could intensify and suppress our experience of emotion [12]. Finally, the "contagion" step refers to our ability to "catch" emotions, that emerge from the interplay from subconscious mimicry and feedback [25]. This approach to emotional contagion has been criticized, as it presents it as a largely automatic bottom-up process when there are times that intentional, explicit attending, as well as appraisal of ones own emotions, can direct and give rise to emotional contagion [2, 18]. In terms of what we observed in our study, we hypothesize that both a mimicry and appraisal-based form of emotional contagion could be occurring. For one, participants had an implicit understanding of the mapping between their face and particle-avatars. This was evidenced by them knowing when others were laughing or expressing other feelings. It is possible that implicit alignment of facial expressions could have been happening even before conscious awareness. However, this wasn't entirely implicit. Many participants reported "wanting" to align their expressions with everyone else in the call. This implied a level of social cognition was occurring which saw them explicitly think about their own behaviour and adjust it to match that of the crowd. Instances of observed emotional alignment was encouraging evidence that the particle visualizations provide sufficient information for the participants to come to emotional conclusions about it.

6.2 Visual Verisimilitude & Emotion Representation

There is an open question of the *verisimilitude* of the visualizations–meaning whether the visualizations themselves genuinely represent the real emotions experienced by the audience, and by extension if audience members accurately perceive those emotions. Co-Here was built assuming a constructed theory of emotion, so a major design goal was to provide a medium that was "sufficiently ambiguous" for users to come to their own conclusions of the emotions of the audience. In other words, utilizing the user to classify and construct their own emotions, instead of the computer system categorically determining what emotions were being experienced. It is thus assumed that the emotions users perceived were not exactly the same, though it would be interesting grounds for future work to explore this question experimentally.

Of note was that discussion around negative valence emotion was underrepresented during focus groups, despite half the presentations having negative valence content. This could be because allowing presentations to be improvised meant that some participants were more playful, which could lend itself to more positive affect presentation styles. Many images from the IAPS that were high arousal and low valence were excluded from the slide show presentations due to explicit content that was unsuitable for day-to-day videoconferencing (e.g., body mutilation). This selection criterion, when applied to the IAPS dataset, could have affected the final quality of negative valence stimuli. Finally, an under-representation of negative valence emotions could also be a shortcoming in the expressive capabilities of the system, where the most detectable emotions skewed positive valence.

6.3 Interactions Between Ethics, Identity and Self Expression

Participants reported that "stress free" aspects of Co-Here enabled them to more freely express themselves. An oft-cited reason for this was the pseudo-anonymity that the visualization provided. There may be a "sweet spot" of anonymity that enables users to express themselves the most. Anonymity provides a disinhibition effect, removing social barriers that may lead to greater self-expression [39]. However, this may likewise lead to increased toxicity and hate, as infamously demonstrated in anonymous online message boards and

games [32, 39]. This calls into question whether the anonymous aspects of Co-Here can be construed as a dark pattern, leading to antisocial behaviour. In the case of online chatrooms, it was shown that anonymity, invisibility, and eye-contact may all contribute to toxicity online [32]. In small groups, Co-Here is semi-anonymous, and semi-invisible. In some cases participants can be identified, such as when a single user audibly laughs and their particle-avatar reacts in turn. In larger groups, identification becomes much more difficult. Co-Here does not mask participant's voices, so does not offer the true anonymity of an internet form. Though there was no active toxicity observed from participants, they all used Co-Here in an experimental setting, and participants could see each other's face before and after using the device. This presents an interesting future design challenge: what is the right level of anonymity to allow free expression from participants, without inviting too much toxicity? Categories 1C - Emotional Expression, 2C - Identity, and 2B - Ethics all intersect with this question. Future work on Co-Here and systems like it could further encourage participants in expressing themselves more by finding the right balance of anonymity.

6.4 System Design Considerations

The system is currently only suitable for small crowds (approximately < 10 people) as with scale the large number of particle-avatars may become overwhelming, and create attentional burdens similar to that experienced while videoconferencing with cameras on. A version of Co-Here for massive audiences would require a new rendering and facial analysis techniques, and would be an ideal iteration to accommodate use cases such as remote concerts or large-scale gaming streams. Co-Here also need not be limited to analyzing the face—the body as a whole offers a rich basis that future versions of Co-Here could use to render affective feedback. The speech signal is also a rich source of affective information and could be further utilized as input.

Co-Here also has a number of technical challenges to solve before it could be used for everyday use. The effectiveness of the landmark tracking used to drive particle avatar animations was prone to noise from low-light conditions. All participants of this study were asked to be in well lit room. More robust landmark tracking is therefore required to use the device in low-light environments. Another technical challenge of the system is that it requires everyone to view their camera and screen straight on, when in reality people's faces are not always guaranteed to be looking towards their camera. For example, people with multiple monitors may appear to be looking to the side since their webcam may not be mounted on the monitor that displays the rest of the callers. A possible solution to this is to use GAN-based facial alignment software that interpolates a set of facial landmarks to a forward-facing mask as a digital puppetry and compression technique [41, 47].

There is much discussion to be had about the fidelity of the particle-avatars used in Co-Here. Prior work investigating realism of avatar forms found that visually low-realism avatars similar to the avatar particles employed in this study elicited inferior reports of copresence and emotion identification compared to a video stream [1]. A raw video stream is still the "gold standard" for getting clear picture of what another user is experiencing. But keeping track of many video feeds can quickly become intractable as the call size grows. The aforementioned experiment was focused on 1:1 calls, not 1:N video calls, where there are unique constraints on screen space and attention. It is an open question as to whether higher fidelity avatars will offer greater affective awareness during calls, and what their corresponding attentional burdens may be, especially when the number of active videocallers grows to encompass tens to hundreds of people.

7 Conclusion

Through in-situ use, we have obtained qualitative evidence that Co-Here gave participants emotional awareness of the audience—this was best demonstrated in observed reports of alignment in emotion (1A-Alignment). Co-Here encouraged emotional expression from users, and provided a platform to give emotional support (1C-Promotion of Emotional Expression). These two findings taken together is suggestive that the system was facilitating affective grounding between users. To establish affective ground, interlocutors must have a shared understanding of eachother's emotions [27]. The reported emotional alignment is evidence that emotions were collectively perceived, understood, and reacted to. Users intentionally manipulated their own expressions as a way emotionally communicate to others, which entails a level of emotional understanding of both other user's emotions, the and how the user's own emotions could be perceived. These are signs that Co-Here affectively grounded the presentations in this study. This claim could be falsified in the following ways: (1) if participants were not cooperative, and trying to deceive others with their emotional expressions, and (2) if the expressions rendered through Co-Here were not accurately perceived. Regardless, the presence of affective grounding isn't a categorical distinction. The finding that there was meaningful awareness of others emotions, and that participants were able to emotionally interact with one another without relying on voice, text, or video, does suggest that Co-Here facilitates the affective grounding of user conversation in some way.

It was discovered that Co-Here was best suited for audience members. For broadcasters, the animations were largely ignored due to the high visual demand of the slide shows. Future work for such a system could render affective feedback in alternate modalities, such as sound or touch, as to assist in the unique attentional demands between broadcasters and viewers as discovered in this study. It is an open question as to what user personalities best suit Co-Here. A presenter who chose to improvise her lines used Co-Here explicitly for audience feedback. However, less confident presenters were very occupied with what slides were coming next and the overall structure of their presentation that they did not look at the visuals. The same can be said of presenters who were less familiar with their presentation material.

Co-Here presents an interaction space between having cameras off and on, encouraging emotional expression from users. It shows promise in its ability to facilitate the communication of affect in a continuous, implicit manner during 1:N video calls, especially for viewers, and further offers a new method to establish affective ground with videoconferencers. The animations provide a supportive environment, enabling participants to understand, share, and align emotions without relying on explicit computational classification of emotional states.

References

[1] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder. The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. *Presence: Teleoperators and Virtual Environments*, 15(4):359–372, 2006.

[2] P. B. Barger and A. A. Grandey. Service with a smile and encounter satisfaction: Emotional contagion and appraisal mechanisms. *Academy of management journal*, 49(6):1229–1238, 2006.

[3] L. F. Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[4] L. F. Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

[5] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[6] A. Begel, J. Tang, S. Andrist, M. Barnett, T. Carbary, P. Choudhury, E. Cutrell, A. Fung, S. Junuzovic, D. McDuff, et al. Lessons learned in designing ai for autistic adults. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–6, 2020.

[7] K. Boehner, R. DePaula, P. Dourish, and P. Sengers. How emotion is made and measured. *International Journal of Human-Computer Studies*, 65(4):275–291, 2007.

[8] M. Bruder, D. Dosmukhambetova, J. Nerb, and A. S. Manstead. Emotional signals in nonverbal interaction: Dyadic facilitation and convergence in expressions, appraisals, and feelings. *Cognition & emotion*, 26(3):480–502, 2012.

[9] P. Bucci, L. Zhang, X. L. Cang, and K. E. MacLean. Is it happy? behavioural and narrative frame complexity impact perceptions of a simple furry robot's emotions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2018.

[10] J. Y. Cho and E.-H. Lee. Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *Qualitative report*, 19(32), 2014.

[11] H. H. Clark and S. E. Brennan. Grounding in communication. 1991.

[12] C. Darwin. The expression of emotions in man and animals. new york: Philosophical library. *Original work published*, 1872.

[13] L. C. De Silva, T. Miyasato, and F. Kishino. Emotion enhanced multimedia meetings using the concept of virtual space teleconferencing. In *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, pp. 28–33. IEEE, 1996.

[14] B. Downe-Wamboldt. Content analysis: method, applications, and issues. *Health care for women international*, 13(3):313–321, 1992.

[15] I. Duboyskii, A. Shabanova, O. Sivchenko, and E. Usina. Architecture of cross-platform videoconferencing system with automatic recognition of user emotions. In *IOP Conference Series: Materials Science and Engineering*, vol. 918, p. 012086. IOP Publishing, 2020.

[16] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

[17] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[18] H. A. Elfenbein. The many faces of emotional contagion: An affective process theory of affective linkage. *Organizational Psychology Review*, 4(4):326–362, 2014.

[19] S. Garrod and M. J. Pickering. Why is conversation so easy? *Trends in cognitive sciences*, 8(1):8–11, 2004.

[20] S. Garrod and M. J. Pickering. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2):292–304, 2009.

[21] J. Guo and S. R. Fussell. A preliminary study of emotional contagion in live streaming. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pp. 263–268, 2020.

[22] J. T. Hancock, K. Gee, K. Ciaccio, and J. M.-H. Lin. I'm sad you're sad: emotional contagion in cmc. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 295–298, 2008.

[23] M. Hassib, S. Schneegass, P. Eiglsperger, N. Henze, A. Schmidt, and F. Alt. Engagemeter: A system for implicit audience engagement sensing using electroencephalography. In *Proceedings of the 2017 Chi conference on human factors in computing systems*, pp. 5114–5119, 2017.

[24] C. Hatfield and J. Cacioppo. Rapson (1994) emotional contagion.

[25] E. Hatfield, J. T. Cacioppo, and R. L. Rapson. Emotional contagion. *Current directions in psychological science*, 2(3):96–100, 1993.

[26] H.-F. Hsieh and S. E. Shannon. Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288, 2005.

[27] M. F. Jung. Affective grounding in human-robot interaction. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 263–273. IEEE, 2017.

[28] S. Kim, M. Billinghurst, G. Lee, M. Norman, W. Huang, and J. He. Sharing emotion by displaying a partner near the gaze point in a telepresence system. In *2019 23rd International Conference in Information Visualization–Part II*, pp. 86–91. IEEE, 2019.

[29] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[30] A. Kristoffersson, S. Coradeschi, and A. Loutfi. Towards evaluation of social robotic telepresence based on measures of social and spatial presence. In *Human Robot Interaction*, 2011.

[31] P. J. Lang, M. M. Bradley, B. N. Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3, 1997.

[32] N. Lapidot-Lefler and A. Barak. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2):434–443, 2012.

[33] C. Latulipe, E. A. Carroll, and D. Lottridge. Love, hate, arousal and engagement: exploring audience responses to performing arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1845–1854, 2011.

[34] E. Lee, J. I. Kang, I. H. Park, J.-J. Kim, and S. K. An. Is a neutral face really evaluated as being emotionally neutral? *Psychiatry research*, 157(1-3):77–85, 2008.

[35] B. B. J. Li and A. Z. Yee. Understanding videoconference fatigue: a systematic review of dimensions, antecedents and theories. *Internet Research*, (ahead-of-print), 2022.

[36] L. Menenti, S. C. Garrod, and M. J. Pickering. Toward a neural basis of interactive alignment in conversation. *Frontiers in human neuroscience*, 6:185, 2012.

[37] P. H. Mui, M. B. Goudbeek, C. Roex, W. Spierts, and M. G. Swerts. Smile mimicry and emotional contagion in audio-visual computer-mediated communication. *Frontiers in Psychology*, 9:2077, 2018.

[38] P. Murali, J. Hernandez, D. McDuff, K. Rowan, J. Suh, and M. Czerwinski. Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2021.

[39] L. Nitschinsk, S. J. Tobin, D. Varley, and E. J. Vanman. Why do people sometimes wear an anonymous mask? motivations for seeking anonymity online. *Personality and Social Psychology Bulletin*, p. 01461672231210465, 2023.

[40] G. M. Olson and J. S. Olson. Distance matters. *Human–computer interaction*, 15(2-3):139–178, 2000.

[41] M. Oquab, P. Stock, D. Haziza, T. Xu, P. Zhang, O. Celebi, Y. Hasson, P. Labatut, B. Bose-Kolanu, T. Peyronel, et al. Low bandwidth videochat compression using deep generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2388–2397, 2021.

[42] M. Pasupathi, L. L. Carstensen, R. W. Levenson, and J. M. Gottman. Responsive listening in long-married couples: A psycholinguistic perspective. *Journal of Nonverbal behavior*, 23(2):173–193, 1999.

[43] A. Ranjan, J. Birnholtz, and R. Balakrishnan. Improving meeting capture by applying television production principles with audio and motion detection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 227–236, 2008.

[44] V. Rivera-Pelayo, J. Munk, V. Zacharias, and S. Braun. Live interest meter: learning from quantified feedback in mass lectures. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 23–27, 2013.

[45] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[46] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.

[47] S. Sharma. Ai can see clearly now: Gans take the jitters out of video calls. *The Official NVIDIA Blog*, 2020.

[48] M. Shermer. Patternicity: Finding meaningful patterns in meaningless noise. *Scientific American*, 299(5):48, 2008.

[49] D. Swarbrick, D. Bosnyak, S. R. Livingstone, J. Bansal, S. Marsh-Rollo, M. H. Woolhouse, and L. J. Trainor. How live music moves us: head movement differences in audiences to live versus recorded music. *Frontiers in psychology*, 9:2682, 2019.

[50] J. L. Tracy and D. Randles. Four models of basic emotions: a review of ekman and cordaro, izard, levenson, and panksepp and watt. *Emotion review*, 3(4):397–405, 2011.

[51] C. Wang and P. Cesar. Physiological measurement on students' engagement in a distributed learning environment. *PhyCS*, 10:0005229101490156, 2015.

[52] C. Wang, E. N. Geelhoed, P. P. Stenton, and P. Cesar. Sensing a live audience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1909–1912, 2014.

[53] R. Wataya, D. Iwai, and K. Sato. Ambient sensing chairs for audience emotion recognition by finding synchrony of body sway. In *The 1st IEEE Global Conference on Consumer Electronics 2012*, pp. 29–33. IEEE, 2012.

[54] K. E. Watkins, A. P. Strafella, and T. Paus. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8):989–994, 2003.