TouchTales: A Care-Centered Protocol for Recognizing Authentic Emotion from Naturalistic Touching and Telling

Rubia Guerra[1], Laura Cang[1], Po-Yu Chen[2], Nao Rojas[3], and Karon E. MacLean[1]

***Abstract*—Naturalistic human touch expression can be an emotionally potent modality, and promising as an informative but unobtrusive input into machine-learning models of affect. However, application-ready emotion-aware technologies must be trained on labeled samples of authentic (felt) emotion of a significant range of intensity and valence. These may come at significant, even traumatic, personal cost for any modality.**

We examined (a) performance of the novel touch modality, and (b) how a care-centered protocol might manage personal burden. Participants (N=5; 3 team members), shared autobiographical stories that elicited powerful emotional dynamics, in 1-3 sessions each (total 10). During storytelling, incidental touch was captured on a pillow-mounted custom flexible 10×10-taxel pressure sensor, alongside physiological signals; then labeled with multiple passes of rich, multimodal self-reports. Protocol, study and analysis design prioritized reflexivity and participant experience.

***Accuracy:* Participant-specific, touch-only models predicted emotion direction (trajectory slope) with 65.2±16.3% accuracy (2s windows; chance 25%, physiology-only models 64.1±16.9%), confirming the value of this unobtrusive channel.**

***Personal cost:* Qualitative analysis contributed an extensive picture of the emotional toll of generating such data, but also some benefits. We offer recommendations for sustainable ethical sourcing of affective data which balance personalization, performance, therapeutic insight and participant care.**

***Index Terms*—Affective Computing, Modeling human emotion, Ethics, Human-computer interaction, Moral implications, Haptics applications.**

I. Introduction

Humans engage with their environment through touch to functionally manipulate it, and also for hedonics and communication with others. Recognition of the importance of specifically *affective* touch has grown in the last decade, in part based on neuroscientific evidence linking touch to our affective systems [1], and the shared experience of widely curtailed interpersonal touch during the COVID-19 pandemic for a developmentally significant period [2].

While some emotion-touch encodings likely exist [3], other evidence suggests that affective meaning is not typically tied to *gesture* (a physical, archetypal movement such as 'stroke' or 'poke') [4]: one gesture can have many meanings, and vice versa. We suspect that affect is more evident in *how* a touch action is expressed [5], [6], as well as the context in which it is interpreted. But, even given this nuance, if humans (and perhaps domesticated animals [7]) can decode emotion from touch alone, could a machine reach a similar degree of reliability? This would open the door to computational models able to interpret affective touch, permitting touch-sensitive devices to respond in emotionally plausible ways.

Setting aside questions of what emotion really is [8], there are several obstacles for collection and modeling protocols. Controlled laboratory settings and artificial stimuli are not ideal for eliciting the dynamic emotion of everyday life. Constrained protocols often focus on a limited set of basic emotions. Self-reporting limited to discrete labels oversimplifies the continuously evolving and contextual nature of emotional experience. Finally, we lack naturalistic datasets (emotionally diverse and grounded in real contexts) to train and compare models which will be robust and practical in real settings. This requires data that captures the nuanced and natural expressions that might occur with a pet or a loved one at home, or an inanimate comfort object (a worry stone or favorite stuffed toy) in a tense setting.

In this work, we explore the feasibility of computationally inferring a dynamically shifting emotion trajectory from naturalistic touch within an emotionally aligned and intense in-lab experience. Because labeling of emotion *state* may discretize the emotion experience too coarsely, we use *emotion direction* to differentiate the experiences of stressed-but-calming-down from stressed-and-getting-more-stressed, represented as an evolving slope of an emotion curve as it moves continuously between Stressed and Relaxed [6]. Our elicitation and labeling protocols are informed by participants' perspectives, for closer alignment with subjective emotional experiences.

We started from Cang *et al.*'s previously validated collection and labeling protocol [6], [9], which draws on cognitive-behavioral (CBT) and dialectical behavior (DBT) therapy in encouraging individuals to lean uninterrupted into their emotional experience. Cognitive interpretation [10] and granular data labeling occur post-task while participants review a recording. Because this affective information is difficult to capture by a single method, we utilized multiple self-report formats to label observational data, and compared models based on touch to those with physiological markers.

We made three main innovations. *Elicitation:* We replaced [6]'s stressful game with reliving a prompted personal experience in a home-like environment created within a lab. This more open-ended and personal task better grounds the elicitation in real life, but with enough control for labeling and analysis rigor.

Free touch as model input: Participants held and stroked a haptically appealing, touch-sensed object while storytelling; [6] senses keypress force during gameplay.

The third relates to *participant experience*. Reliving an intense experience can be hard on all involved. Along with

*This work was supported in part by Natural Sciences and Engineering Council of Canada (NSERC).

[1]Rubia Guerra, Laura Cang, and Karon MacLean are with Faculty of Science, Department of Computer Science, University of British Columbia, Vancouver, Canada. {rubiarg,cang,maclean}@cs.ubc.ca

[2]Po-Yu Chen is with the Faculty of Arts & Science, Department of Computer Science, University of Toronto, Toronto, Canada.

[3]Nao Rojas is with the Faculty of Science, Department of Computing Science, University of Alberta, Edmonton, Canada.

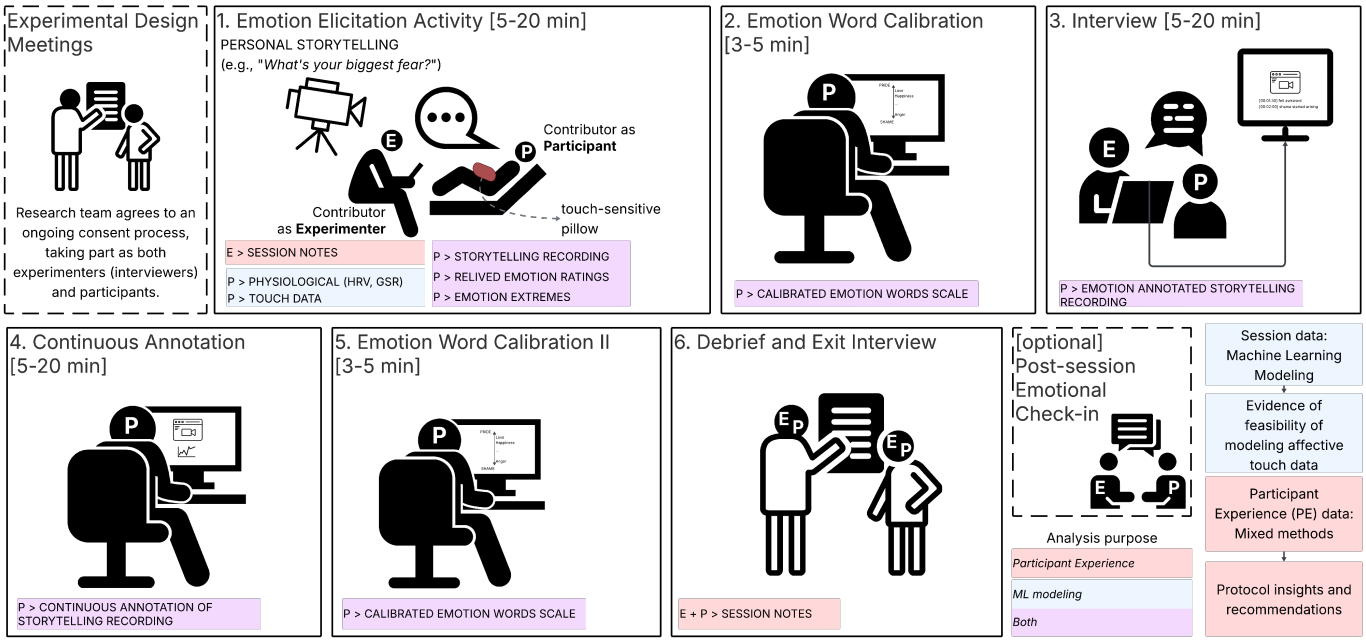


Figure 1: **Overview of Study Protocol** balancing care-centered participation with capturing rich, model-ready emotion data. Individuals contributed model and participant experience data in two roles, conducted within a comfortable, safe study space.As *experimenters,* they fostered a safe, participant-led conversation with ongoing consent. In other sessions, some of the same contributors were *participants*, telling personal stories then labeling the video in multiple calibration and annotation passes.

confirming the presence of affect information in this modality, we need to make collection feasible before we can hope to generate population data from a larger group. Can the process be completed in a timely manner without undue stress? Do our protocols accurately represent the emotions felt; can a model capture the subtleties of emotional expression in a way that is both meaningful, and eventually deployable in real-world affective systems?

We collected 95 minutes of storytelling for analysis: high-commitment, emotion-rich touch, video and physiology (Figure 1). To verify affective content in this modality, we trained participant-individualized machine learning (ML) models from touch and physiological data, then examined both performance of models trained on multiple, emotionally diverse sessions, and individual differences in touch expression. We contribute:

1) An initial *performance benchmark for models classifying emotion direction in touch expression* during emotional storytelling, based on 5 participants (9 analyzed sessions).
2) A *proof-of-concept of a training data collection protocol and its evaluation*, prioritizing consent, participant comfort, and rich emotion labeling. Through qualitative analysis of participant and researcher experience and workload, we derived insights and recommendations into the feasibility and effectiveness of collection and training.

II. Related Work

Researchers have assessed emotion state with many modalities and methods: self-reports, facial expression and physiological signals offer varied lenses. We review the intersection of touch, emotion, and computational modeling to ground the potential of modeling affect through touch.

A. Touch, Emotion and Modeling their Connection

Touching can induce emotional responses, with skin mechanoreceptors involved in moderating touch perception and its emotional sequelae [1], [11]. In nonglabrous skin, Pacinian corpuscles are sensitive to deep pressure and vibrations, and CT afferents to gentle stroking and associated particularly with positive emotions, such as pleasure and comfort [12]. A-fibers, meanwhile, are involved in transmitting pain and fine tactile discrimination. Integration of these signals contributes to a complex experience which influences emotional regulation, stress reduction, and social bonding.

Recent studies indicate that CT afferents are also present in glabrous skin previously thought to lack them (palms and fingertips — the body part that usually *does* the touching) [11], [12]. This satisfyingly aligns with our lived experience (we respond emotionally to touching nice or unpleasant things, as well as to being touched), and supports searching for a touch*er's* emotion signifiers in their tactile behavior.

Behaviorally, touch has been related to a range of emotive information, from comfort/soothed to alarmed/distressed [3], [13]. Since people can often identify emotions based on touch alone [12], [1], touch combined with ML techniques holds promise for emotion estimation. Indeed, touch has recently been used to classify emotions accurately from physical interactions, such as keystrokes or haptic feedback [14].

However, this behavior is complex and individual, as are touchers' concordant emotion responses [15], [16]. We need to

verify that the patterns are individually repeatable enough to be machine-recognizable. Meanwhile, touch models require large training datasets, must interpret complex cues (*e.g.,* finger strokes, pressure patterns), and (as for any modality processed in real-time) address hardware and computational demands. This work helps to define and chip away at these obstacles.

B. Enhancing Ecological Validity in Emotion Labeling

Emotion is a complex and dynamic process, shaped by social, psychological, and physiological factors [8]. Traditional emotion labels, often oversimplified and detached from real-world contexts, risk missing the nuances and fluctuations of genuine emotional experiences [17]. Approaches commonly used to label emotional data, including forms of experience sampling [18], [19] (*e.g.,* daily diaries, ecological momentary assessment) may not be feasible for fine-grained collection, particularly in naturalistic settings, and may still suffer from some degree of recall bias due to reporting delay.

A different approach captures the ongoing dynamics of emotional states by first curating an immersed, recorded experience, followed by multiple passes of high-granularity emotion self-report in different formats based on retrospectively reviewed evidence [20], [6]. The recorded media reduces recall bias, and the result is a fine-grained view of emotion evolution linked to environmental conditions.

We novelly examined the participant's experience in a protocol of this kind in a homelike setting, to induce nuanced and authentic emotion trajectories (experienced, rather than simulated or imagined), which unfold naturally in a conversational interaction, mediated by a comforting and physically-inviting prop that evoked a pet or a treasured object.

C. Emotion Elicitation Through Personal Storytelling

In emotion research, participants are often asked to revisit charged memories to evoke specific responses [21], inducing authentic emotion and robust engagement [22], [23]. Relatedly, clinical practices often involve cognitive restructuring and guided imagery, where therapists help patients reflect on emotional experiences to promote emotional awareness, processing and a more positive reappraisal. We similarly encouraged participants to engage with their memories and emotional states in a safe setting. Our conversational approach was inspired by therapeutic techniques that focus on the bodily basis of experiences. Relevant examples include the Focusing Method [24], and techniques from Dialectical Behaviour Therapy (DBT) [25], which similarly emphasize an objective, non-judgmental approach to present-moment awareness.

This reflective and therapeutic process helps participants remain engaged while interacting with the system, and provides a framework for later interpretation. In post-review, we asked participants to reiterate their momentary response throughout the recording, and also encouraged them to reflect on how resolved their experience felt in relation to the original event, in hopes that these elements would jointly foster a framework for cognitive interpretation over time.

D. Physiological Signals as Emotion Proxies

Physiological signals such as galvanic skin response (GSR) and heart rate variability (HRV) are often used as proxies for emotional states, offering valuable insights into the autonomic nervous system's reaction to emotional stimuli. GSR is sensitive to arousal [26]. HRV correlates with stress, anxiety, and relaxation, and can be an indicator of the individual's emotion regulatory activity [27], [28].

We triangulated physiological signals with self-reports for a broader picture and more robust models, and as a baseline for assessing the relative contributions of touch data *vs.* more-studied physiological and self-report measures.

E. Computational Models of Affect

ML algorithms can be used to analyze and interpret touch-based emotion data. These computational models aim to recognize, predict, and classify emotional states from diverse inputs. Common approaches include rule-based methods and deep learning models (patterns learned from data without explicit programming [14]). Deep learning can find complex patterns in large datasets, achieving high classification accuracy from images, text, and physiological signals. However, its black-box form obscures insight, *e.g.,* regarding pattern mappings or how individuals differ.

Regardless of type, most published affect recognition models identify only static states, with less attention to *dynamic emotion evolution*: the continuous shift in emotional experience over time [29]. This gap is relevant for any modality seeking to assess realtime, evolving experience.

F. Researchers-as-Participants for Elicitation Insight

Self-reporting is a crucial tool in emotion research, but limitations (oversimplification of emotional states, poor reproducibility, context dependence, constrained ecological validity [8]) have driven interest in complementary techniques like physiological sensing and observational coding [30]. However, little attention has been paid to how participants themselves conceptualize and articulate emotional labels that are both meaningful to their lived experience and computationally actionable in ML systems. This is where first-person research methods offer a valuable approach [31], emphasizing the researcher's own subjective, bodily, and situated experience as a source of knowledge.

Experimenters themselves can be a valuable resource here: actively engaging in our own studies offers first-hand insights into both the investigated phenomena and the participant experience [32]. This enhances collection rigor, data quality and phenomenological robustness [33], [32], [34]. Some elicitation studies have employed and illustrate benefits of an experimenter-as-participant approach [35], [36], [37], [38], [39]. For the Stanford Emotional Narratives Dataset (SENDv1), researchers annotated emotional narratives [39]. Though not participants in the traditional sense, researchers' immersion in annotation deepened their sensitivity to the temporal and contextual dynamics of emotion. This has also helped pioneer measurement techniques; *e.g.,* comparing physiological indicators with self-reported emotions revealed

discrepancies best understood via first-hand experience [40]. Micro-phenomenology, for example, is a qualitative research method designed to attend to, articulate, and analyze lived experience in fine detail, which is particularly relevant for understanding embodied and temporal experiences [41].

However, there is little about data collection for computational emotion modeling, raising concerns about taking these experiences for granted. We target this gap by positioning experimenters-as-participants to systematically document the tensions between subjective emotional granularity ("*Did I feel betrayed or just disappointed?*"), and the less nuanced categorical labels required by standard emotion recognition pipelines. Our goal is to grasp the participant experience and learn to maintain its safety, while learning enough about emotion dynamics to generate emotion data suitable for computational modeling and recognition protocols.

III. Methods

We collected biometric and touch data from a primary emotion task of emotional storytelling, then generated self-reported dynamic emotion labels by adapting [20]'s multi-pass self-report labeling protocol. Each session has six stages (Figure 1). The initial storytelling task was the most sensitive, with high-emotion, instrumented data collection; the rest entailed different forms of review and self-reporting on the storytelling experience. To glimpse individuals' expressive range and nuance, we mixed single with multiple emotionally diverse sessions per participant.

A. Setup

The study utilized a custom touch-sensitive object, physiological sensors, and a data collection interface (Figure 2).

Setting: To put participants at ease and encourage the naturalistic storytelling central to our study, we furnished the study room with a couch and other simple décor and arranged it to resemble a comfortable living space, avoiding the sterility and formality that often marks in-lab experimental settings. A photograph of the setting in the supplemental materials gives additional context (see S.1-Fig.6). Participants were alerted to the video-camera; an opaque divider maintained some privacy and focus.

Touch Object (Sensed Pillow): We used a custom flexible, fabric-based[1] touch sensor previously described and validated for capturing social touch gestures [42], here a 10x10 taxel grid on a curved pillow. It captures multiple simultaneous touches (multitouch) in 100-taxel frames, with each taxel's pressure scaled to [0-1023]. Like [43], [44], it detects 5g–1kg on a 10"×10" surface (one taxel/inch^2). Fingerpad-size taxels are sufficient for emotion tasks, which generally incite broad rather than precise movements [3], [44], [45]. Resolution is balanced with computational efficiency, size, cost and deformability.

[1]Built from commercially available piezoresistive and conductive fabric. Fabric is commercially available at www.eeonyx.com.

Physiology: We used a Grove GSR sensor[2] for analog skin conductance, with adjustable sensitivity. An optical Pulse Sensor Amped[3] detected blood volume changes and thence heartrate (HRV), with amplification and noise cancellation.

Study Interface: We developed a custom tool to support *session administration*; *live sensor visualization* for researcher monitoring of data streams in realtime; participants' *emotion word calibration* task, ranking emotion words on a vertical axis; researcher *interview annotation* with timestamped comments; and participant *joystick annotation* for continuously rating of their emotional state while reviewing their storytelling video.

Sampling and Synchronization: All data streams (touch sensor, GSR, pulse, and interface inputs) were sampled at 54Hz with an Arduino Mega microprocessor[4], synchronized and timestamped for consistent temporal alignment.

B. Study Support Team (*SST*) and Roles

The *SST* consisted of four co-authors who interacted with participants and/or accessed non-aggregated, identifiable data. The 1st and 2nd authors were *SST co-lead*s. The *SST* pooled diverse experiences, with backgrounds in engineering and computer science, aged 20-40 years, and cultures from throughout Asia and North and South America. None were trained psychologists. Three *SST* members were also participants, contributing to the dataset alongside external participants (§II-F). Therefore, study roles rotated during sessions, as detailed in Table I. The experimenter who processed the dataset and ran modeling protocols did not act as a participant. No *SST* participant saw their own raw data during analysis, and they did not model their own machine learning analysis. They only accessed aggregated results and discussions of them. For the reflexive thematic analysis, all *SST* members were involved in reflections and discussions.

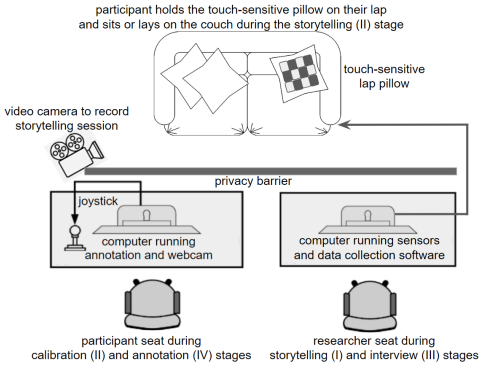
***SST* Role**	**E1**	**E2**	**E3**	**E4**	**Note**
Study design	x	x	x	x	
Participant (***Part***)	x	x	x	-	Experimenters contributing data.
Interviewer (***Int***)	x	x	-	-	Senior *SST* member, in-room.
Technical support	-	x	x	x	Setup, monitoring; in+out-of-room.
Protocol reflection	x	x	x	x	
ML analysis	-	-	-	x	Modeling & analysis; handled identifiable data. Junior *SST* member w/ senior supervision; out-of-room.
PE analysis	x	x	-	-	All members involved in discussions. Senior researchers conducted mixed-methods analysis on participant experience data. Out-of-room.

Table I: **Roles taken by the four *SST* Experimenters**. Two experimenters were in room with *Part* in each session.

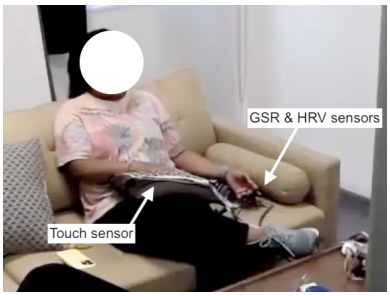
[2]Seeed Studio. *Grove - GSR Sensor*. https://wiki.seeedstudio.com/Grove-GSR_Sensor/. Accessed March 6, 2025.

[3]World Famous Electronics LLC. "Pulse Sensor Amped." https://pulsesensor.com/pages/pulse-sensor-amped-arduino-v1dot1. Accessed March 6, 2025.

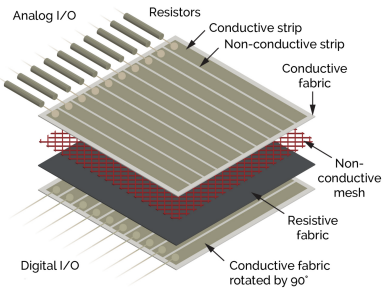
[4]Arduino Mega 2560 Rev3, accessed April 8, 2025, https://docs.arduino.cc/hardware/mega-2560



(a) Physical setup: room layout and relative positioning for participant and researcher, for all stages of data collection.



(b) Photo of the participant setup.



(c) Fabric pressure sensor schematic [42].

Figure 2: Overview of the experimental setup, showing layout and participant positioning.

Beyond standard user-study methodologies, ethics, and technical knowledge, *SST* preparation involved discussion and evaluation after each data collection session. These reviews were devoted to ongoing ethical soundness, methodological rigor, and sensitivity to participants' experiences. For example, the fact that experimenters and external participants had pre-existing personal connections helped establish trust and rapport during sessions, but also necessitated conscious effort to maintain objectivity during data collection and analysis. As an exploratory study, we sought to minimize bias through a number of mechanisms (§VII also summarizes concerns and improvements that influenced study procedures over time):

Workshopped questions and protocol with multiple external experts: Feedback on our proposed protocol from four research groups with psychology expertise resulted in three distinct protocol iterations, which addressed subtle points of bias and improved prompt wording.

Used open-ended emotion probes with emotionally neutral language: We were careful to avoid questions that inferred participant feelings (*e.g.,* "you seem anxious"); instead, our script includes neutral prompts (*e.g.,* "how do you feel?") to encourage participants to use their own terms.

Separated data processing from data generation and collection: The experimenter who processed the dataset and ran the machine learning protocols did not act as a participant, creating an additional layer of separation.

C. Piloting and Prompt Design

We piloted the full protocol with two *SST* participants, iterating on the consent process clarity and prompt quality. To help participants identify strong emotional stories, the *SST* brainstormed an initial prompt set by following theories of autobiographical memory and emotional recall commonly used in emotion research [46], [21]. In pilots, we refined the set for evoking both positive and negative emotions, and balance in emotional intensity and resonance, and safety and accessibility. Final prompts are listed in Table II.

D. Recruitment, Safety and Ongoing Consent

Inclusion & Recruitment: Due to the sensitive nature of personal stories, we included only participants who expressed that they (1) were comfortable sharing experiences with one of the researchers as they would when venting or sharing with a friend; (2) could answer frankly in a post-session check-in about their ability to manage any trauma, should it be triggered. We found we were best able to meet this criteria by recruiting within the research team's social circles, including from the *SST*. We targeted our outreach to these social circles with a recruitment form.

We discussed *SST* member involvement in a dedicated meeting, explicitly considering potential power dynamics between senior and junior experimenters/participants to ensure junior members did not feel pressured into roles they were uncomfortable with. Similar sensitivity to this issue was extended to participants from the external recruitment pool.

When a potential participant (including *SST* members) indicated interest, a *co-lead* explained the protocol and detailed which data would be reviewed for analysis and by whom. The *co-lead* also helped establish the participant's privacy, including their comfort with specific *SST* members reviewing their data. Participants could indicate if they preferred certain team members only have access to anonymized versions, rather than transcripts or video recordings.

Initial Consent: Participants (*Part*) were given up to a week to consider participation following the initial interview. Initial consent was obtained with a conventional process. Then, at the start of each session, *Part* indicated or renewed consent by signing a paper form or assenting online.

Ongoing Consent: Due to the storytelling task's triggering potential, we used several tactics to maintain safety. (1) An explicit ongoing mechanism wherein the interviewer (*Int*) would ask only twice about any item; then, if *Part* was not forthcoming, move to a different topic or line of questioning. (2) The establishment of a 'safe' word or phrase to indicate the need to move on from a topic. (3) "Anything goes" encouragement with respect to expression, including cussing, yelling, singing, over- or under-sharing, *i.e.*, whatever would allow

Table II: **Storytelling prompts and the most prominent emotions elicited** as reported by participants. Each prompt was used once. The **first** Reported Emotion in each row is the one the participant reported as predominant across the story. Reported 'Opposites' refer to what the participant believes to be the opposite emotional experience to the emotions they reported actually experiencing. Mean (SD) duration of all stories was 9:30 (SD 4:18) minutes.

Prompt (positively biased)	**Reported Emotions**	**Opposites**	**Duration (min)**
What are you most proud of?	**Pride**, Contentment	Shame	4:48
Tell me about the person you share news with first.	**Love**, Pride, Gratitude	Sadness, Loss	8:53
What is the nicest thing that's been done for you?	**Pride**, Accomplished	Doubt, Helplessness	5:57
When did you feel the most satisfied with your life?	**Connectedness**, Longing	Loneliness, Anger	7:12
What do you remember about the best period of your life?	**Excitement**, Nostalgia	Shame, Embarrassment	7:00
Prompt (negatively biased)		**Mean (SD)**	**6:46 (1:31)**
What is your biggest fear?	**Anxiety**, Dread	Accomplished, Fulfilled	10:55
What is the biggest stressor in your most important relationship?	**Guilt**, Sadness	Fulfillment, Satisfaction	13:15
What was the hardest decision you've ever made?	**Anxiety**, Fear, Gratitude	Disgust	7:43
What is your biggest frustration?	**Anxiety**, Confusion	Satisfaction, Gratitude	9:40
What is your current biggest worry?	**Regret**, Longing	**Anger**, Sadness, Spite	19:33
		Mean (SD)	**12:13 (4:33)**

them to authentically experience their feelings in a safe and comfortable manner. (4) Monitoring for signs of distress, such as prolonged silence or visible discomfort, and responding with a check-in and offer of a break, switch to a lighter prompt, or early session end. (5) After the storytelling task, a chance to decompress in a manner they were comfortable with. (6) An optional check-in scheduled 12-48 hours later. The consent negotiation process added 2-5 min to the data collection session, depending on whether the participants had more questions or comments.

E. Data Collection & Labeling Protocol

Each session consisted of initial consent plus six study stages, taking 30-85 minutes in all. Table III summarizes the resultant data and abbreviations.

1) Storytelling (5-20m): We requested *Part* to sit or lie on the couch, placed HRV and GSR sensors ($\boldsymbol{Bio}_d$) on their non-dominant hand, and asked them to keep their dominant hand on the touch-sensitive pillow ($\boldsymbol{Touch}_d$) located on their lap or chest, depending on their body posture. During sessions, we noted that participants tended to keep the pillow stationary on their lap or against their chest, with most touch interactions occurring with their other hand; in part this was due to their wearing sensors on one of their hands.

Prompt counterbalancing minimized tone bias. Within a session, prompts were randomized for order effects. Across sessions by a given participant, the first prompt's valence was alternated to balance initial emotional tone.

We administered prompts (Table II) flexibly to prioritize participant comfort. *Part* could skip any prompt they found unproductive or distressing without explanation, and *Int* would offer an alternative from the same valence category; if *Part* seemed distraught, *Int* offered to pause the session, allowing them to regroup. Other safety and ongoing consent mechanisms (above) were maintained.

As *Part* responded to prompts, *Int* asked clarifying questions and probed for emotional content and deeper emotional memories, in a dialogic rather than rigidly structured manner.

Table III: **Data Produced** for each session. † indicates data_d items (in contrast to labels_l) used in the present analysis; others are listed for completeness, used in analyses reported elsewhere. Researcher field notes were collected in all stages.

Stage / Items	**Description**
1. Storytelling:	*All streams synchronized on one time-axis*
· Video & Audio stream	Of participant (whole body)
· Biometric stream	†$\boldsymbol{Bio}_d$: HRV & GSR, sampled at 54Hz
· Touch pressure stream	†$\boldsymbol{Touch}_d$: TP, 10x10 taxels/frame, 54Hz
· Relived Emotion Ratings	†$\boldsymbol{RelivEmoRat}_d$: Custom scale for emotion's *Similarity* to original, *Intensity* & *Resolution*
2. Calibration:	*Categorical emotion representation*
· Calibrated Words	$\boldsymbol{CalWords}_d$: Participant word choices (from defined list) & placement within $CWScale_d$
· Calibrated Word Scale	†$\boldsymbol{CWScale}_d$: Participant-specific extrema
3. Interview: · Discrete Annotations	Participants' phrases and comment annotations aligned on *Storytelling* time axis
4. Continuous Annotation: · CA time-series	†$\boldsymbol{CntAnn}_d$: Aligned on *Storytelling* time axis, and *Calibration* emotion scale (y-axis)
5. Exit: · Post-Session Debriefing	†***Participant reflections*** on protocol effectiveness and emotional processing
· Recalibrated Words	$\boldsymbol{CalWords}_d$ (updated from *Calibration*).

These clarifying questions were used to guide *Part* toward more detailed emotional descriptions and to manage the scope of the stories. For example, once a strong emotional reaction (*e.g.,* crying or uncontrollable laughter) emerged, *Int* would steer the conversation toward a closure on that specific topic, enabling a detailed, fine-grained account of the moment. The pre-existing relationships and established trust between the *Int* and the *Part* facilitated this process, allowing deeper exploration of emotionally charged topics.

After the storytelling was completed, we removed the physiological sensors and touch-sensitive pillow. *Part* moved to a computer station for the session's remainder.

We then asked *Part* to reflect on three questions about the

emotions that they had just felt (*Relived Emotion Ratings, or* ***RelivEmoRat***$_d$*)*. On a 10-point Likert scale, they rated the emotion *Similarity* to the original occurrence; *Intensity* compared to the original occurrence; and *Resolution* of the events or emotions at present.

2) Calibration (3-5min): To frame the emotional landscape of their story, we asked *Part* to name the most prominent emotion they had just experienced; then to contextualize with words on a linear, graphical scale. The words they chose here could also prime the participants to employ them or their common synonyms in later stages. We chose this approach because we needed an instrument capable of representing greater complexity than a single valence or value, following what past participants have directly informed us of their emotion experience.

*Scale Creation (****CWScale***$_d$*):* A scale was displayed vertically on a graphical user interface with drag/drop functionality (see S.2-Figure 7 in the supplementary materials). *Int* labeled the extrema with *Part*'s named emotion, then asked *Part* for an opposite emotion word for the other end, if needed suggesting words until *Part* was satisfied; *e.g.,* if the most-prominent term was "Caring", the opposite end might be "Indifference". Finally, *Part* was consulted about which of these words should sit at the top *vs.* bottom of the axis.

To keep the scale personally meaningful, *Part* was free to choose any word that felt most relevant to their story. If they had difficulty, *Int* collaboratively brainstormed options, drawing from the list described in the next step.

*Scale Population (****CalWords***$_d$*):* From a pre-set list of 12 emotion words (*fear, love, sadness, happiness, disgust, surprise, embarrassment, envy, pride, sympathy, gratitude*, and *anger*), *Part* was instructed to drag as many of the provided 12 emotion words as they felt were relevant (8-12 suggested) onto a vertical scale, positioning each based on its perceived relationship to the extrema they had just defined. For cross-session consistency, in this step *Part* was constrained to the provided list. By the end of this step, the *CWScale*$_d$ had become a participant-defined subjective scale that captured the nuances of that individual's reported emotional experience, beyond categorical labels; and having the capacity to represent mixed-valence emotions.

3) Interview (5-20min): For detailed insight into how participants experienced their emotions during storytelling, we asked them to revisit the audio-video recording of their storytelling. For a timeline consistent with the original emotion expression, *Int* operated a custom interface that allowed them to run and pause the video while annotating *Part*'s reflections on its timestamps. *Part* discussed the recording with *Int*, giving particular attention to emotionally poignant moments, explanations for their expressions or verbalizations, unusual or unexpected behaviors (*e.g.,* sudden laughter in an otherwise sad story), sudden breaks in prose, *etc.*

This stage was conducted informally, often extending the story beyond the recording's scope as *Int* asked for or *Part* volunteered rich interpretive details and descriptions of their unfolding emotional states.

4) Continuous Annotation (5-20min): Part re-watched the Storytelling video all the way through without pausing. As they watched, they manually annotated their emotion experience on the Calibration stage emotion scale (*CWScale*$_d$) using a custom unbiased 1-axis joystick [47], producing a corresponding continuous trace with a vertical position for every data point in the stream (***CntAnn***$_d$; Figure 3). *Part* was instructed that the joystick's 'up' position (away from themselves, top position on the produced screen trace) corresponded to the top of the emotion scale, and its 'down' position to the opposite emotion they had named. When held still, the joystick produced a horizontal line, representing a steady emotion state.

5) Exit (<30m): The session closed with three final steps.

Recalibration: To assess stability of the emotion landscape over the session, *Part* repeated the "population" step of the previous Calibration stage. Specifically, they dragged the same set of emotion words they had previously chosen, described in the first calibration step onto a blank vertical scale labeled with the previously provided extrema. Participants were instructed to focus on recalibrating the ordering and positioning of the set of words, starting from a blank state to avoid biasing the recalibration process.

Interview: In an open-ended conversation, *Int* obtained ***Participant Reflections***, *e.g.,* about *Part*'s feelings throughout the session, intensity of emotion recall, and their interpretation of the emotion scale across labeling stages. *Int* (and *Part* if also an experimenter) took notes.

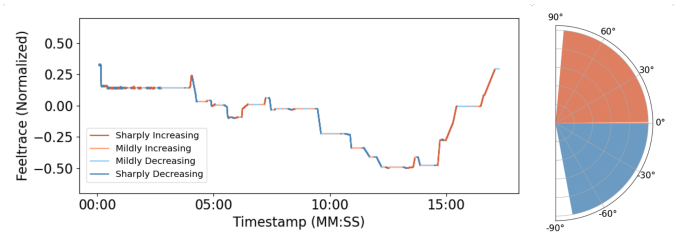
Well-Being Check: Int conducted a final well-being check to ensure *Part* felt stable and comfortable before leaving. This included confirming their emotional state, revisiting their earlier data access permissions to ensure they remained comfortable with them, and verifying that we could follow up with them. *Int* scheduled the follow-up based on *Part*'s preference, typically 12-48 hours post-session.

IV. Machine Learning Models

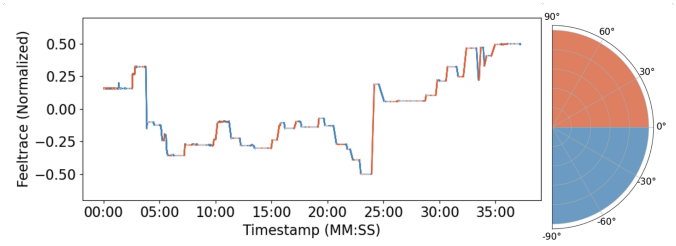
We required our ML framework to provide insight into feature importance, optimal estimator choice, parametrization, resulting performance data (confusion matrices and scores), and robustness (defined here as model ability to generate consistent results across multiple data subsets).

This dataset's high dimensionality (multiple data streams), temporal nature, and personalized, subjective emotion labels influenced our model choices. These characteristics required us to prioritize approaches that include feature selection, dimensionality reduction, time series analysis, personalized modeling, data synchronization, and interpolation.

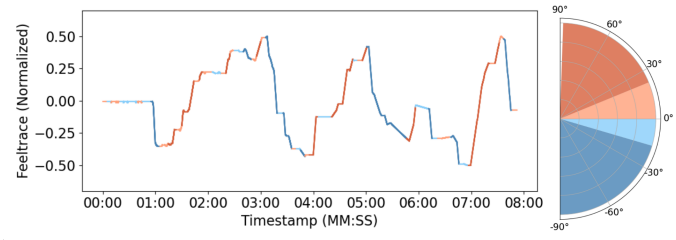
The principal output of our trained, personalized models is prediction of emotion direction (specifically, the ***EmoDir***$_l$ label), chosen because it most elegantly captures emotion dynamics. It is trained on the slope of participants' continuous emotion annotation (*CntAnn*$_d$), as detailed below. Model classification accuracy is based on ability to predict discrete bins of *EmoDir*$_l$ values derived from *CntAnn*$_d$.



(a) Example 1



(b) Example 2



(c) Example 3

Figure 3: Feeltrace position trajectory (right) and $EmoDir_l$ (slope) bin distributions (left) for real 'example' data (participant number not disclosed for anonymity). The right panel displays the time series of the joystick's position, colored by its corresponding slope bin. The left panel shows the polar distribution of these Feeltrace trajectories, demonstrating the distribution of reported emotional directions.

A. Time-Series Preprocessing

Our modeling input features were sourced from sampled time-series. Their preparation primarily consisted of filtering; we excluded samples without valid $CntAnn_d$ annotations.

Bio$_d$ *&* ***Touch***$_d$ *Streams:* To reduce instrument noise, we applied a 1D polynomial Savitzky-Golay filter on both GSR and HRV, originally sampled at 54 Hz. We used a window length of 100 samples (for both 2s and 5s) and a polynomial order of 1, for effective smoothing while preserving important features. $Touch_d$'s 10x10 taxel frames were smoothed by calculating the frame-wise sum across all taxels, then applying the same 1D Savitzky-Golay filter. We note that this frame-wise sum does not differentiate between a high-force contact with a small portion of surface area and a low-force contact with a large portion of surface area, a limitation which we hope to mitigate in future senor technology.

CntAnn$_d$ *Stream:* To address joystick jitter due to hand tremor and/or sampling artifacts, we used an exponential weighted moving average (EWMA) model. Through trial and error with visual inspection, we found that a decay parameter of 0.3 balanced dynamic responsiveness with noise reduction. Finally, we linearly interpolated missing values and time-aligned them to the video clock.

Partitioning of Streams into Windows: We partitioned all time series (Bio_d, $Touch_d$, $CntAnn_d$) into equal-sized windows prior to modeling. Based on prior work in emotion and touch analysis [6], we used two window lengths: 5s for broader emotion context, and 2s for finer temporal resolution. We experimented with intermediate window sizes in 0.5s increments but report on only these two sizes for simplicity.

Stream Merging: We used nearest-neighbor timestamp joins with a 1s tolerance, aligning timestamps to the same origin.

Table IV: Input features for each model schema.

Schema	Temporal Features	Frequency Features
Physiological (GSR, BPM) 39 features	Mean, var, max, min, area under curve (AUC), sum of abs differences (SAD)	Power spectral density in delta, theta, alpha, beta, and gamma bands
Touch 6 features	Mean, var, max, min, AUC, SAD, calculated over large-area touch (LAT: sum of touch values/window)	*Note: Touch-frequency features were excluded from final models due to lack of performance gain*
Combined 45 features	All from both physio and touch schemas	All from both physio schemas

B. Input Feature and Training Label Definition

1) Input Features: Informed by Cang *et al.* [6]'s findings on touch pressure's relation to emotion expression, we extracted input features from the preprocessed time series for three modeling schemas (Table IV) to capture participants' (i) biosignal fluctuations, (ii) touch behaviors, and (iii) both.

*2) Training Labels (****EmoDir***$_l$*):* This analysis relies on a single label type, $EmoDir_l$, defined by drawing on the interpretation for emotion as directional and changing ("where I am" *vs.* "where I'm going") [20]. We derived labels by computing $CntAnn_d$'s slope, then assigning to bins. Here, we used three bins to capture generally *increasing, neutral* and *decreasing* slope. Figure 3 depicts this process.

Because the annotations tend to concentrate around the zero-slope region, and distributions exhibited individual variance, we allocated slope data into equally-sized bins, with participant-specific boundaries. Figures 3's bin polar distributions illustrate this mild variance.

To create this balanced distribution, we computed slope for sliding $CntAnn_d$ windows and transformed values using the arctangent function, which maps values onto a bounded angular scale between -90° and +90°. These extrema represent the joystick moving instantaneously (vertical negative or positive slope) but not into the past. We divided this 180° range into four quantile-based bins (roughly equal numbers of samples/bin) by sorting the slope values and calculating percentile thresholds (0%, 25%, *etc.*) using NumPy's[5] percentile function. Finally, slope values were binned using the 'np.digitize' function.

[5]NumPy: https://numpy.org

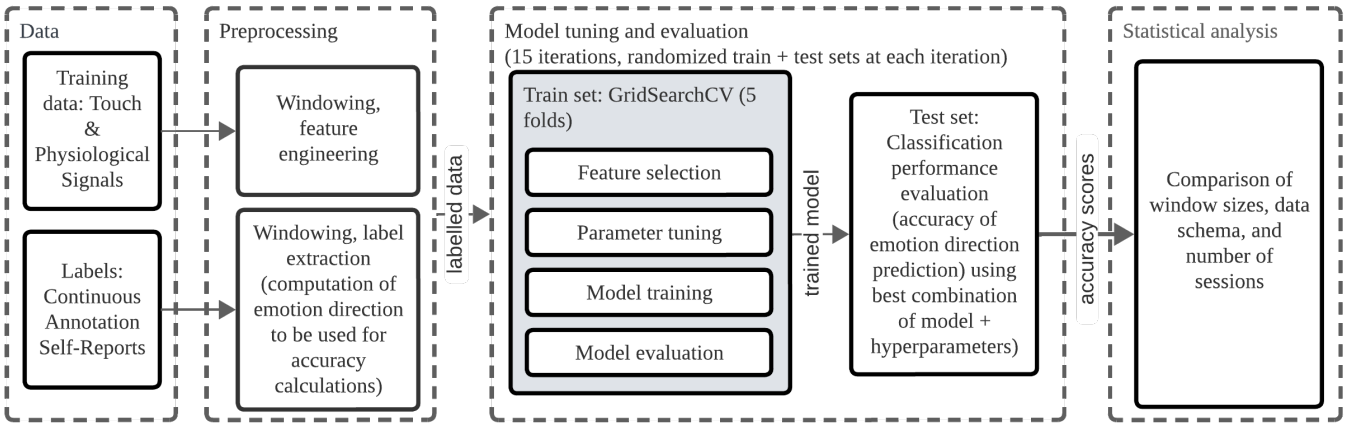


Figure 4: **Pipeline for model tuning and evaluation.** We performed grid search cross-validation ($k = 5$) on the training set to tune hyper-parameters and select best-fit models for the source data. Models were then evaluated on a new test set to calculate performance metrics. We repeated this process 15 times per participant-session, and report mean test scores across the 15 runs and 9 sessions.

This proof of concept analysis used only $CntAnn_d$ data, balancing the capture of emotional dynamics with simplicity. We note that it is possible to extract other labels, following [47], and this could likely increase accuracy with minimal additional computational effort.

C. Model Selection & Training

We employed four distinct classification algorithms (Extra Trees, Random Forest, AdaBoost Classifier, and Gradient Boosting), all sourced from the Scikit-Learn ensemble machine learning library [48]. To optimize model performance, we determined the number of estimators (*i.e.*, individual trees or models within an ensemble) to be either one or two multiples of the number of features. Specifically, if the dataset comprises N features, our preliminary number of estimators was set to N and 2*N. For the Gradient Boosting Classifier, we explored learning rates of 0.8 and 1.0.

The training process involved constructing a Scikit-Learn Pipeline by integrating the RFECV and GridSearchCV classes. RFECV, from the Scikit-Learn feature selection library, executed recursive feature elimination with cross-validation to identify optimal features, reducing dimensionality and potential overfitting. GridSearchCV, from Scikit-Learn's model selection library, systematically explored parameter values for each estimator.

Training with 5-fold cross-validation, we collected feature importance, confusion matrices, performance scores, and the best parameters associated with the most effective estimator. This helped us to select robust models with parameter configurations optimized for the specified input features.

V. Results & Discussion: Modeling Performance

Five unique participants, 20-40 years old, contributed 10 sessions: two external participants each contributed one, and three *SST* members either two or three (Table V). All were previously known to at least one *SST* member. Other demographics are omitted due to the small pool.

For one session, at the participant's request we discarded all but the post-session reflections, which we still incorporated into our experience commentary (§VI), leaving 9 full sessions for modeling and quantitative analysis. This was this participant's only session.

These sessions provided four interdependent data sources (Table III): two quantitative (synchronized streams, quantitative ratings), two qualitative (post-session participant debriefings, researcher field notes.) The streams directly supported modeling and correlation analyses, covered in the present section. The remaining sources are analyzed in §VI.

We evaluated individualized model performance by comparing $EmoDir_l$ prediction accuracy across three factors: input modality (Modality_f), session number (SessN_f), and window size (WinSize_f). These factors respectively capture the relative information in touch versus physiological data, the impact of emotion-type diversity across sessions, and the effect of temporal aggregation on computational efficiency.

A. Model Evaluation

We trained four ensemble classifiers (Extra Trees, Random Forest, AdaBoost, Gradient Boosting) using a 5-fold grid-search CV pipeline (Figure 5), repeated 15x/participant-session. Personalized models yielded a **mean top accuracy of 65.2% ± 16.3%** (touch-only at 2s windows, chance 25%). Figure 5 shows accuracy distribution by (a) session, (b) modality, and (c) window size.

This result is reasonably comparable to prior work [6], which found F1-scores of up to 0.82 (82%) to classify emotion evolution from a different touch modality. While [6]'s accuracy is higher, we note that these studies have importantly different study conditions with the present study's presenting a greater classification challenge. In [6], participants engaged with a stress-inducing video game, and touch was measured as keypress force for navigation keys. In addition to a distinct touch modality, that task was more narrowly defined and likely yielded a more homogeneous emotional signal that is easier

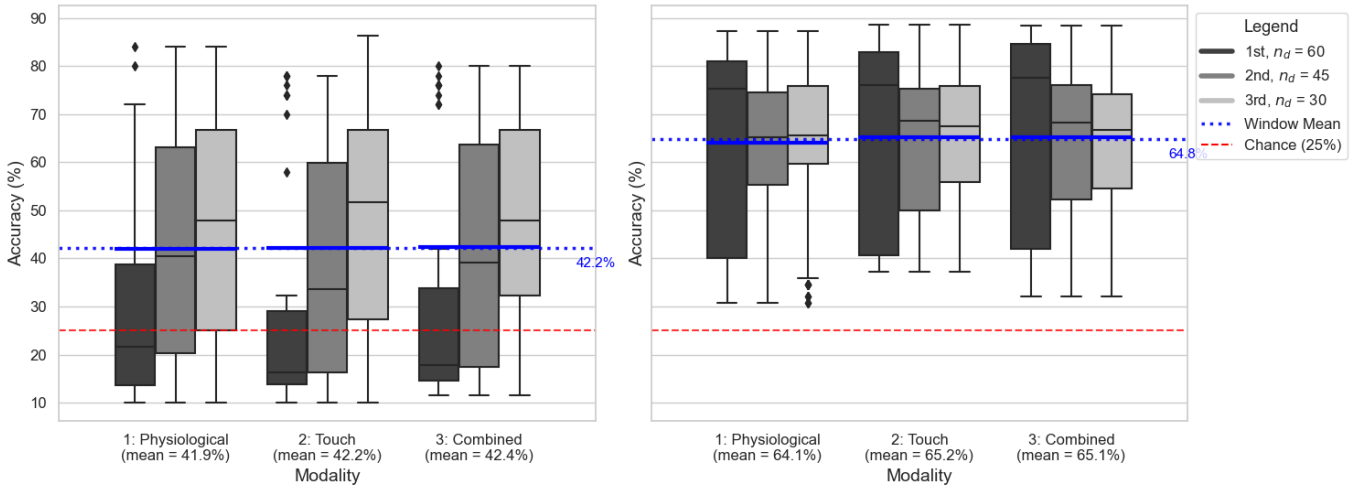


Figure 5: **Model accuracy in predicting *EmoDir*$_l$ (chance = 25%) by analysis factors of WinSize$_f$, Modality$_f$, and SessN$_f$.** Boxplots show percent accuracy for three input modality schemas (physiological, touch, or both) across up to three sessions at two window sizes (5000 and 2000ms). In legend, 1$^{\text{st}}$ *etc.* means SessN$_f$, and *e.g.,* $n_d = 60$ to # of accuracy scores for that modality input in a full session (15 accuracy scores each for N=4 unique participants).

to classify. In contrast, the present study employs open-ended prompts that elicit a broad spectrum of emotional experiences within a naturalistic, conversational setting, introducing greater linguistic and temporal variability and making the classification problem intrinsically more challenging.

B. Statistical Analysis

We tested the effects of WinSize$_f$, SessN$_f$, and Modality$_f$ on accuracy by combining accuracy results from ML models trained individually per participant into a single dataset. This combined data was analyzed using a Generalized Mixed Linear Model (GMLM) with REML estimation and a random intercept for the entire test, fitted using the Powell optimizer.

1) Assumption checking: Violations motivated use of a GMLM with REML, which is robust to non-normality and unbalanced group sizes. For residuals, Shapiro–Wilk on each factor returned $p < .001$, indicating violations of univariate normality. For variance homogeneity, Levene's tests showed unequal variances for SessN$_f$ and WinSize$_f$ ($p < .01$), but not for Modality$_f$ ($p > .05$).

2) GMLM Regression Analysis: Regression results are analyzed below (full results in S.4-Tables VI and VII). The model uses 810 observations from 4 participants[6] (15 to 45 observations / participant across all input modalities, per window, per input modality), clustered by participant ID.

Baseline Accuracy (Intercept): The intercept of **0.651** (SE 0.065, $z = 10.061$, $p < .001$) represents the predicted accuracy when Modality$_f$=physiological only, WinSize$_f$=5000 ms, and SessN$_f$=1 session.

Session Effect: Coefficient = -0.018, $z = -0.387$, $p = 0.699$. This number is the expected change in accuracy when moving from one session to the next *for the same window size and modality*. However, the effect is not statistically significant, suggesting that the addition of emotion-type diversity across sessions does not impact model accuracy. While the marginal averages in Figure 5 show Session 3 with higher raw accuracy, when window size and modality are controlled the session-to-session slope is not significant.

Window-Size Effect: Coefficient = -0.359, $z = -8.885$, $p < .001$. Larger windows (5000ms) incur a statistically significant drop in accuracy relative to 2000ms windows, indicating a clear effect of window size on model performance, with a negative impact on accuracy.

Modality Effect: Coefficient = 0.019, $z = 0.469$, $p = 0.639$. No significant difference between physiological, touch, or combined inputs: adding touch signals did not boost accuracy in a meaningful way.

Interaction Effects: No two- or three-way interactions reached significance ($p > .5$ for all but session *vs.* window size), indicating that session, window size, and modality do not interact to affect the accuracy.

Participant Variance: The random-intercept variance ($\sigma_u^2 = 0.010$) is small compared to fixed effects, implying limited between-participant heterogeneity in overall accuracy, a desirable outcome given our use of individualized models.

C. Generalized vs. Individualized Models

Using a leave-one-out retraining pipeline (Figure 5), we compared generalized (trained on N–1 participants) to individualized models. The latter significantly outperformed generalized models ($\rho = 0.950$, $p < .001$). In generalized models, window size had a small positive effect on accuracy ($\rho = 0.000$, $p < .001$), and the two-way interaction between model type and window was significant ($p < .001$). No modality effects or higher-order interactions emerged.

[6]One participant excluded from ML analysis, as per their request.

VI. Results & Discussion: Collection Methods

To assess and refine the collection protocol for participant comfort, ethical considerations and researcher workload, we analyzed the qualitative data components and quantitative Relived Emotion Ratings (Table V).

The SST conducted a reflexive thematic analysis (RTA) on rich qualitative data from researcher field notes and participant debriefing sessions. We immersed ourselves through iterative coding and reflecting on patterns to identify high-level themes. Following [49], we considered the active role of the researchers' perspectives in data interpretion, positioning our subjectivity as a valuable analytical instrument, to capture nuanced insights into participant experiences and procedural dynamics. We refined themes for coherence and depth through recursive review, documenting how our assumptions and data interactions shaped this development.

A. Consent Process

We discussed with participants how the comprehensive ongoing consent process (which we estimate added 2-5min to session length for the initial discussion, and variable time for periodic check-ins during the session) aligned with the strain that they actually experienced, with respect to either gaps or excess. While the initial consent presentation was brief, the constant check-ins were the real burden, implicitly conveying an expectation from the researcher that the participant was experiencing distress, which could potentially interfere with the emotion elicitation process or distract from the emotional task itself.

Participants: Generally felt comfortable with the protocol and appreciated the concern for their safety. Foreknowledge about data collection and usage may have empowered the participant who asked to delete a recording post-session. Three participants suggested more efficiency, particularly through shorter consent check-ins, finding the repetition somewhat distracting and potentially reinforcing the idea that their stories were inherently distressing.

Researchers: Found the process's repetitiveness demanding. The increased session length due to extensive check-ins contributed to their fatigue. However, they also reported that its thoroughness increased their confidence that participants felt safe, comfortable and empowered to set their own boundaries, and to some extent this validated their effort.

Recommendation: Moderately reduce conservatism by (a) retaining upfront highlighting of data recording and planned usage, with explicit reminders of participant options at session ends; and (b) streamlining by limiting check-ins to brief ones pre- and post-session.

B. Emotions in Storytelling

Participants experienced a broad range of emotions in response to prompts (Table II), often simultaneously or in conflicting valence combinations. This complexity highlights the potential richness of storytelling as an elicitation method.

Participants: Lives are personal and varied, and prompts elicited many distinct emotions, intertwined in complex ways; thus the prompted emotions were not always predictable. For every prompt (session), participants named more than one emotion, sometimes of contrasting valence. For example, when asked to relive the "hardest decision you've ever made", P2 reported feeling Anxiety, Fear, and Gratitude as the strongest or most prominent relived emotions, and acknowledged their mixed valence. Others listed Longing alongside Connectedness, and Nostalgia next to Excitement.

Researchers: Found that this response variability, particularly when mixed in valence, later complicated cross-session annotation, increasing analysis burden and possibly threatening model convergence.

Recommendation: So-called "inconsistency" will often be real, but our models cannot yet handle it. For improved modeling effectiveness while we learn, future protocols could focus on smaller emotion sets known to be reliably elicited (*e.g.,* Anxiety or Pride) across many participants. This would simplify annotation and improve model accuracy through more consistent data.

C. Experiencing Emotion Evolution

During sessions, we listened as participants' memories triggered strong responses: they cried, paused to collect themselves, and made connections to past or current events. We tried to give space for these feelings to resolve.

Participants: Described sessions as cathartic, experiencing emotional release through storytelling, and appreciated having time to resolve strong feelings.

Researchers: Also found sessions emotionally taxing due to empathetic engagement with participants' intense feelings.

Potential Intrinsic Benefit for Participants: Many stories involved personal reflections reminiscent of think-aloud journaling. Research has shown significant benefits to cognitive and emotional processing via journaling [50], [51], whether written or spoken [52]. It may be that confronting or inviting strong negative feelings can introduce a co-activation such that positive feelings are experienced in the aftermath [53].

For technology intended to help with addressing and/or attending to negative feelings, we wonder if such a protocol of expressing and examining strong feelings may have a two-fold benefit: to (1) emulate the natural progression through negative and positive emotions for training data; and (2) provide a guided process to help release existing emotional tension [50] in consultation with clinical experts.

Recommendation: Additional research is needed to confirm the potential intrinsic benefit in an emotion-capture collection protocol of providing emotional processing opportunities, and refine it by *e.g.,* explicitly integrating clinical expertise to oversee safe resolution. Regardless, shorter sessions or fewer prompts per session could reduce participant and researcher burden, making repeated sessions more feasible.

D. Duration in Emotion Stories

Negatively valenced stories lasted nearly twice as long as positively valenced ones (6.46 *vs.* 12.13 minutes; Table II).

Table V: **Relived Emotion Ratings (*RelivEmoRat*$_d$)** by participant and session (*e.g.,* P3-1 indicates P3's 1st session), and Pearson correlations between dimensions.

P#–Session	**Similarity** *1: not at all 10: perfect match*	**Intensity** *1: not at all 10: perfect match*	**Resolution** *1: very active 10: inactive*
P1-1	4.0	5.0	4.5
P2-1	7.0	8.0	2.0
P3-1	8.5	4.0	3.0
P3-2	6.0	6.5	2.5
P4-1	10.0	9.0	1.0
P4-2	8.0	6.0	8.0
P4-3	10.0	9.0	1.0
P5-1	3.0	7.0	2.5
P5-2	2.0	2.0	8.0
P5-3	6.0	4.0	1.5
Mean (SD):	**6.85 (2.69)**	**5.95 (2.39)**	**3.95 (3.36)**
Pearson's r:	*Sim-Int:* 0.63	*Sim-Res:* -0.48	*Int-Res:* -0.61

Participants: Reported that their frequent pauses during negative stories were to manage strong emotions, or to introspect before providing detailed context for their feelings.

Researchers: Observed that longer negative stories required greater attentiveness and emotional labor, increasing fatigue.

Recommendation: To balance experimental cost and data quality, future protocols might lean towards negative prompts that tend to resolve toward calm states, reducing burden while still eliciting data that can model emotion evolution.

E. Relived Emotion Ratings and Memory Resolution

Participants' ratings demonstrated complex interactions between the similarity, intensity, and resolution of their relived emotions to those of the original experience (Table V). Stories involving ongoing emotional conflicts or recurrent feelings elicited particularly strong ratings as well as phenomenological responses identified in qualitative analysis of their debrief interviews.

Participants: Compared to resolved narratives, retelling unresolved stories (*e.g.,* P2's career dilemma, P4-3's childhood memory) generated higher similarity (r=0.632) and intensity ratings relative to original events. Paradoxically, these accounts showed weaker resolution correlations ($r = -0.475$), consistent with the retelling surfacing active emotion processing. Some participants volunteered that articulating childhood-era feelings might help reframe but not fully resolve them.

Researchers: Facilitating unresolved narratives required heightened ethical vigilance. Stronger signals ($r = -0.611$, intensity-resolution correlation) seemed to come at the cost of increased participant distress management. Childhood-origin stories resurfaced present-day emotions, complicating temporal anchoring of affective states. Finally, their mixed valence required continuous dialogue to disambiguate "then *vs.* now" emotional layers.

Recommendations: To balance the modeling potential of unresolved emotions with participant agency, future work could integrate participant-guided somatic reflection (*e.g.,* "How does this story feel in your body right now?") to ground experiences, respect narrative control and provide temporal labeling anchors. In parallel, whether a computational model can detect chronic *vs.* situational affective patterns could be determined by pairing resolved and unresolved story variants (same participant, different time-points).

These speculative strategies aim to balance ethical safeguards with the scientific value of capturing in-process emotional reactivation (a phenomenon particularly interesting for technologies addressing longitudinal emotional growth).

F. Ambiguity in Continuous Annotation Scale

Our findings support the potential of using continuous input to capture subjective data which is inherently dynamic, such as the 1-dimensional graphical-physical joystick used here, an underexplored approach [54]. We could not have attempted this analysis using traditional discrete-point reporting surveys such as the PANAS [55] and Discrete Emotions Questionnaire [56]. However, refinement is needed.

Participants: Generally interpreted the scale as intensity-based but occasionally struggled with ambiguity about whether it represented intensity or valence, and used varied annotation strategies. They described it as "*how intensely did I feel [the emotion at the top]*" (P3; similar to P2, P4); or "*more like a binary rather than a spectrum [and] felt like it was a little bit bouncing across a binary*" (P1). In contrast, P5 may have treated it as valence: "*could have been more like generically positive or negative feelings*"'; but like P1, P5 also indicated that they "*head[ed] to the end of the scale for stronger emotions [and] stayed middling otherwise*".

Physical annotation strategies ranged from moving the joystick "*stepwise*" (P5) to "*crank[ing] up or down for extreme emotions*" but "*reset[ting] to neutralish*" (P4). P1 noted that they were "*playing with the range at first*" before "*reliv[ing] the experience to feel the matching trace*".

Researchers: Observed inconsistent interpretation of scale usage among participants, the attempted addressing of which then complicated data analysis and model training efforts.

Recommendation: Explicitly instructing participants on scale interpretation (intensity *vs.* valence) with a brief training period seems effective in enhancing modeling accuracy and reducing confusion, but needs to be further developed. The graphical input employed here allowed participants to express nuanced and personalized emotional complexity and continuity. The result can be displayed back to the user for immediate feedback in the form of a dynamic visualization. This kind of continuous input linked to participant-driven visualization could support richer, more accurate data for emotion research.

G. Synthesis: Mitigating the Cost of a Powerful Tool

The interplay between the ecological validity of storytelling as an elicitation device and its methodological complexities is a central tension in this collection approach. We confirmed past evidence that storytelling has a substantial capacity to capture lived emotional reactivation [9], based on observed correlations between emotional similarity, intensity, and resolution. While we do not have a basis for direct comparison to other

approaches, our closest case is [6]'s gameplay elicitation (conducted by the same authors), where emotions were authentic rather than simulated by the participant; but were observed to be far less intense than here, and had little personal meaning or relevance to real life.

However, the elicitation fidelity and power of this tool impose a cost, counted in annotation consistency, participant-researcher burden, and ethical safeguards that scale poorly compared to conventional affective computing paradigms. We offer three important points of guidance for incorporating storytelling into emotion-aware model-building.

1. We need more powerful and less laborious methods for temporal grounding. Multiple passes over the storytelling experience increase reflection opportunities and allow for multi-modal data and labelling capture. However, each pass is cognitively demanding and time-consuming for both the participant and analysis team with diminishing returns for each additional pass. Particularly in the context of personal storytelling, continuous annotation tools must evolve beyond intensity/valence ambiguity to capture the process of emotional reactivation (*e.g.,* distinguishing remembered *vs.* emergent affect).

2. A strong, ethical infrastructure is critical to gain participant trust. Our current tiered consent approach is less efficient than one-time, blanket methods; however, it fostered trust and psychological safety, which in turn likely improved both the quality of the data and the ecological validity of the models we developed, *i.e.*, led to better data. Moving forward, the challenge is to preserve a trust-building structure in methods that can scale to broader, longitudinal or field use.

3. Emotional labor benefits from reciprocity, and may offer therapeutic value. We observed that recounting unresolved experiences was particularly challenging for participants, often emotionally intense, yet described as accompanied by a sense of release. While we do not yet know if this intensity improves modeling quality, it might have value in a more therapeutic parallel purpose. The observed cathartic benefits suggest future protocols might reframe data collection as mutualistic interaction, where the interviewer's nonjudgmental attentive involvement contributes to the participant feeling 'safe', and thereby unburdening themselves. Researchers' data collection labor directly informs adaptive emotional support mechanisms during storytelling.

These considerations set the stage for evaluating the cost-benefit evaluation of naturalistic emotion data collection, seeking a balance between phenomenological depth and sustainable implementation that we analyze next.

VII. Overall Assessment

We consider the implications of these results for building responsive and interactive agents based on dynamic computational emotion models derived from and responsive to touch input, eventually in realtime. Our primary immediate objectives were to establish feasible *modeling* objectives and benchmarks, and design a protocol that supports meaningful and sustainable participant experiences. We reflect on the process of generating training datasets that capture spontaneously evolving emotions, the various costs of doing so, and what would be needed for such models to have practical utility for emotion-aware interactive systems.

Notably, models built on naturalistic, unstructured touch sensed during personal emotion experiences performed comparably to those using physiological signals. Although relative modality performance analyses was not our primary aim, this substantively extends [6]'s related finding for keypress force during stressful game play (also naturalistic but more structured and impersonal), and reinforces the promise and accessibility to emotion inference of the touch modality in diverse forms and contexts.

A. Current Protocol Capabilities and Limitations

The protocol evaluated here, which updates and combines a personal-storytelling paradigm borrowed from [9] and with [6]'s multipass labeling process as well as new elements, demonstrates promise in capturing nuanced, personally grounded emotion dynamics through naturalistic touch. While this establishes a foundation for computational modeling of evolving emotions, it is based on a relatively small dataset which also exhibits variability in emotion expression and annotation interpretation across sessions and participants. The protocol also imposes significant demands on both participants (emotional introspection) and researchers (emotional engagement, lengthy sessions, logistical strain).

In Section VI-G, we outlined possible approaches to *reduce* this burden (*e.g.,* by leveraging profiled generalized models as individual starting points, or developing more intuitive, self-guided annotation tools for participants) or identifying its *intrinsic value* (balancing participant psychological benefit with model utility).

B. Cost-Benefit of Emotional Labor of Data Contributors

While a generalized emotion model targeting a broad population could in principle be created from conventionally compensated contributors, our results show that it will be less accurate than an individualized one. The latter entails significant personal investment in the contribution of emotion data, but the contributor is also the primary beneficiary (increased accuracy). This tradeoff underlies the approach's sustainability: is accuracy (a future utility) sufficient motivation? *Temporal discounting*, a well-studied psychological phenomenon, describes a human tendency to discount future outcomes relative to nearer-term benefits [57], including in choices related to one's health and wellness [58].

Here, there may be an additional, more immediate and intrinsic *therapeutic* benefit. Our participants reported that the storytelling offered cathartic value in a manner akin to journaling. Indeed, there are documented benefits to cognitive and emotional processing via such externalization [52], [50]. This protocol's structured nature potentially provides a guided process for emotional tension release.

To extend the value of generalized models, there is a possibility that most individuals align at least roughly with one of a limited number of "emotion archetypes". In a practical implementation, then, a user might start with a generalized

model created for their closest archetype(s). Their own data contributions would progressively refine its initial functionality, in a rewarding cycle where immediate cathartic benefits also motivate continued engagement that improves long-term model utility. Eventually they accumulate an asset able to respond to their unique emotion patterns through tailored emotion-aware applications.

Ethical and practical sustainability concerns might thus be adequately addressed: emotional labor is motivated in the short term by intrinsic and immediate value, while the improved utility of sophisticated models in applications that serve a need provides longer-term, discounted value.

C. Improving the Protocol

This protocol successfully captured rich emotional data and accurately labeled it; however, participant and researcher burden undermine its scalability. We have identified ways to reduce burden while improving accuracy, marking each by considerations of **E**ffectiveness, **C**ost to participants and experimenters, **D**ata sufficiency, and **F**easibility.

1) Target a smaller, consistently elicitable set of emotions to increase data density and annotation reliability. (E, D)
2) Streamline consent processes, reducing fatigue and overhead by explicitly agreeing up-front on data-recording details and minimizing repetitive check-ins, with low opt-out barriers. (C, F)
3) Provide explicit instructions and training to participants on annotation scales to ensure consistent interpretation and reduce inter-annotator variability. (E)
4) Limit negative prompts to those that reliably resolve toward calm states, enabling shorter sessions that reduce emotional labor but are still emotionally rich. (E, C)

We expect that these adjustments will significantly reduce participant and researcher burdens while maintaining or improving data quality for effective modeling.

D. Defining "Good Enough" For Practical Use

"Good enough" accuracy will be application-dependent, ranging from coarse emotion direction inference for casual support to fine-grained, high-confidence recognition for clinical decision-making. Individuals who cannot communicate emotions verbally may get value just from reliable valence direction detection to enable basic emotional communication and appropriate caregiver responses. Similarly, apps prompting users to journal about their experiences benefit from detecting general emotional shifts rather than precise categories, as the journaling process itself provides therapeutic value. In interactive gaming contexts, detecting broad emotional states such as stress or relief can enhance game dynamics and player experience without requiring granular emotion recognition, making moderate accuracy acceptable for adaptive gameplay. In contrast, ER therapy in a clinical setting might require substantially higher accuracy with more precise emotion categorization to provide reliable intervention recommendations.

Our current $\sim$65.2% $\pm$ 16.3% (touch-only at 2s windows) accuracy with individualized models for predicting emotion direction (§V-A) might be enough for at least the lower end of this utility range. We anticipate that this can be improved with additional training data.

Additionally, prior work (*e.g.,* [47]) demonstrates that models trained with multiple types of labels (*e.g.,* continuous annotation and discrete emotion categories) can better capture the multidimensional nature of emotional experience. Including such labels could help disambiguate trajectories, particularly in mixed or transitional emotional states, potentially boosting prediction accuracy with marginal added computational cost. That said, this would likely require adapting our current modeling pipeline and possibly targeting different use cases. For example, multi-label models may be more appropriate for contexts requiring fine-grained emotion tracking, whereas our current approach offers a lightweight, directionally informative signal. We believe that future work should explore these tradeoffs more systematically, particularly in relation to model generalizability and real-time deployment constraints.

VIII. Conclusions and Limitations

Our findings underscore the potential of touch as a significant indicator of internal emotion shifts, with a modeling approach based on the dynamic and individualistic nature of human emotional expression. This study is the first training procedure of its kind, collecting and labeling data for computational models of naturalistic affective touch that adapt as the user adds to them. By involving a small number of participants in multiple sessions, we were able to explore the feasibility of intermittent and repetitive model training on strong emotional expressions, identifying obstacles to a care-focused and sustainable data collection protocol at each stage. While our participants do not represent the broader population, their insights can guide a more extensive study.

We collected this data in a lab setting. To facilitate responses from touch-centric emotion-aware interactive agents, a data labeling procedure should enable continuous, unsupervised data collection, with accuracy improving as contextually diverse data is integrated. Like predictive text systems on smartphones, such a system must function reasonably well from the outset, ideally featuring a seamless training and data collection protocol. We similarly imagine a live system that offers immediate functionality and refines performance through repeated user interactions and periodic "recalibration" requests. Future work needs to consider classification performance in longitudinal studies, for deeper insights into model evolution and data integration over time.

Finally, we highlight the secondary and preliminary but exciting finding that models built on naturalistic, unstructured touch performed as well as those drawing on more conventional (and less accessible) physiological input. Combined with [6]'s evidence that typing touch pressure outperforms EEG, this bolsters the proposition that touch is both an emotionally informative and technically accessible modality.

Looking ahead, integrating such models into affective, emotionally interactive touch applications holds promise for personal companions, therapeutic aids, and social robots that respond adaptively to users' emotional states. Importantly, our

results suggest that for certain populations, *e.g.,* individuals who cannot communicate emotions verbally, detecting the *direction* of emotional valence (positive or negative) may be "good enough" for providing meaningful, supportive feedback. This approach prioritizes actionable insights and inclusivity, especially in contexts where granular emotion classification is unnecessary or infeasible.

Responsible deployment demands ethical considerations, trust, accountability, and long-term engagement strategies. As emotion-aware technologies become commonplace, we need ways to safeguard user privacy, ensure transparency in data use, and design for sustained, positive user engagement.

In summary, this work lays the groundwork for adaptive, touch-based emotion recognition systems that can evolve with users and support a wide range of applications, from casual companions to clinical aids; and calls us to prioritize ethical, user-centered design in their ongoing development.

ACKNOWLEDGEMENT

Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding this work. Human user research was conducted under UBC Ethics #H15-02611. We thank the contributions SPIN Lab members for their support, particularly Preeti Vyas, Devyani McLaren, Bereket Guta, and visiting researcher, Mirella Hladký. We also thank Prof. Rebecca Todd, Prof. Amori Mikami, and Dr. Katelynn Boerner, all who provided insightful feedback in the early stages of our work.

REFERENCES

[1] J. H. Kryklywy, P. Vyas, K. E. Maclean, and R. M. Todd, "Characterizing affiliative touch in humans and its role in advancing haptic design," *Annals of the New York Academy of Sciences*, vol. 1528, no. 1, pp. 29–41, 2023.

[2] M. H. Burleson, N. A. Roberts, A. A. Munson, C. J. Duncan, A. K. Randall, T. Ha, S. Sioni, and K. D. Mickelson, "Feeling the absence of touch: Distancing, distress, regulation, and relationships in the context of covid-19," *Journal of Social and Personal Relationships*, vol. 39, no. 1, pp. 56–79, 2022.

[3] M. J. Hertenstein *et al.*, "Touch communicates distinct emotions." *Emotion*, vol. 6, no. 3, p. 528, 2006.

[4] S. Yohanan and K. E. MacLean, "The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature," *Int'l J of Social Robotics*, vol. 4, no. 2, pp. 163–180, 2012.

[5] K. Altun and K. E. MacLean, "Recognizing affect in human touch of a robot," *Pattern Recognition Letters*, vol. 66, no. November, p. 31–40, 2014.

[6] X. L. Cang, R. R. Guerra, B. Guta, P. Bucci, L. Rodgers, H. Mah, Q. Feng, A. Agrawal, and K. E. MacLean, "Feeling (key)pressed: Implicit touch pressure bests brain activity in modelling emotion dynamics in the space between stressed and relaxed," *IEEE Transactions on Haptics*, pp. 1–8, 2023.

[7] K. T. Konecki *et al.*, "Touching and gesture exchange as an element of emotional bond construction. application of visual sociology in the research on interaction between humans and animals," in *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, vol. 9, no. 3, 2008.

[8] L. F. Barrett, "The theory of constructed emotion: an active inference account of interoception and categorization," *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[9] X. L. Cang *et al.*, "Discerning affect from touch and gaze during interaction with a robot pet," *IEEE Trans on Affective Computing*, vol. Early Access, no. 01, pp. 1–1, 2021.

[10] A. Uusberg, B. Ford, H. Uusberg, and J. J. Gross, "Reappraising reappraisal: An expanded view," *Cognition and Emotion*, vol. 37, no. 3, pp. 357–370, 2023.

[11] L. K. Case, N. Madian, M. V. McCall, M. L. Bradson, J. Liljencrantz, B. Goldstein, V. J. Alasha, and M. S. Zimmerman, "Aβ-ct affective touch: Touch pleasantness ratings for gentle stroking and deep pressure exhibit dependence on a-fibers," *Eneuro*, vol. 10, no. 5, 2023.

[12] F. McGlone, J. Wessberg, and H. Olausson, "Discriminative and Affective Touch: Sensing and Feeling," *Neuron*, vol. 82, no. 4, pp. 737–755, May 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0896627314003870

[13] M. C. Rozendaal and H. N. Schifferstein, "Pleasantness in bodily experience: A phenomenological inquiry," *International journal of design*, vol. 4, no. 2, pp. 55–63, 2010.

[14] A. Alam, S. Urooj, and A. Q. Ansari, "Human emotion recognition models using machine learning techniques," in *2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON)*. IEEE, 2023, pp. 329–334.

[15] T. Kinnunen and M. Kolehmainen, "Touch and affect: Analysing the archive of touch biographies," *Body & Society*, vol. 25(1), 2019.

[16] M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner, "The communication of emotion via touch." *Emotion*, vol. 9, no. 4, p. 566, 2009.

[17] P. H. Bucci, X. L. Cang, H. Mah, L. Rodgers, and K. E. MacLean, "Real emotions don't stand still: Toward ecologically viable representation of affective interaction," in *2019 8th Int'l Conf on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 1–7.

[18] K. Hoemann, Z. Khan, M. J. Feldman, C. Nielson, M. Devlin, J. Dy, L. F. Barrett, J. B. Wormwood, and K. S. Quigley, "Context-aware experience sampling reveals the scale of variation in affective experience," *Scientific reports*, vol. 10, no. 1, pp. 1–16, 2020.

[19] S. Schneider, D. U. Junghaenel, T. Gutsche, H. W. Mak, and A. A. Stone, "Comparability of emotion dynamics derived from ecological momentary assessments, daily diaries, and the day reconstruction method: Observational study," *Journal of Medical Internet Research*, vol. 22, no. 9, p. e19201, 2020.

[20] X. L. Cang *et al.*, "Choose or fuse: Enriching data views with multi-label emotion dynamics," in *IEEE 10th Int'l Conf on Affective Computing & Intelligent Interaction (ACII)*, 2022.

[21] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment.* Oxford University Press, 2007.

[22] A. Selwood, C. B. Harris, A. J. Barnier, and J. Sutton, "Effects of collaboration on the qualities of autobiographical recall in strangers, friends, and siblings: Both remembering partner and communication processes matter," *Memory*, vol. 28, no. 3, pp. 399–416, 2020.

[23] S. Ozawa, "Emotions induced by recalling memories about interpersonal stress," *Frontiers in Psychology*, vol. 12, p. 618676, 2021.

[24] M. N. Hendricks, "Focusing-oriented experiential psychotherapy: How to do it," *American journal of psychotherapy*, vol. 61, no. 3, pp. 271–284, 2007.

[25] C. J. Robins and M. Z. Rosenthal, "Dialectical behavior therapy," *Acceptance and mindfulness in cognitive behavior therapy: Understanding and applying the new therapies*, pp. 164–192, 2011.

[26] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[27] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding," *Review of general psychology*, vol. 10, no. 3, pp. 229–240, 2006.

[28] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability," *Frontiers in Psychology*, vol. 5, 2014. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01040

[29] P. Bucci *et al.*, "Real emotions don't stand still: Toward ecologically viable representation of affective interaction," in *IEEE Int'l Conf on Affective Computing & Intelligent Interaction (ACII)*, 2019, pp. 1–7.

[30] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[31] A. Desjardins, O. Tomico, A. Lucero, M. E. Cecchinato, and C. Neustaedter, "Introduction to the special issue on first-person methods in hci," pp. 1–12, 2021.

[32] T. Beekman, "Stepping inside: On participant experience and bodily presence in the held," *Journal of Education*, vol. 168, no. 3, pp. 39–45, 1986.

[33] B. Dennis, "Understanding participant experiences: Reflections of a novice research participant," *International Journal of Qualitative Methods*, vol. 13, pp. 395 – 410, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:142814089

[34] F. Friberg and J. Öhlén, "Reflective exploration of beekman's participant experience," *Qualitative Health Research*, vol. 20, pp. 273 – 280, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:29253293
[35] C. Anderson and M. Henry, ""listen and let it flow": A researcher and participant reflect on the qualitative research experience." *Qualitative report*, vol. 25, no. 5, 2020.
[36] L. S. Whiting, J. Petty, B. Littlechild, and S. Rogers, "Undertaking pre-pilot work to gain an empathetic insight into participants' perspectives," *Nurse Researcher*, vol. 29, no. 4, 2021.
[37] R. Johnson, Y. Rogers, J. Van Der Linden, and N. Bianchi-Berthouze, "Being in the thick of in-the-wild studies: the challenges and insights of researcher participation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1135–1144.
[38] E. Wainwright, E. Marandet, and S. Rizvi, "The body–space relations of research (ed) on bodies: The experiences of becoming participant researchers," *Area*, vol. 50, no. 2, pp. 283–290, 2018.
[39] D. C. Ong, Z. Wu, T. Zhi-Xuan, M. Reddan, I. Kahhale, A. Mattek, and J. Zaki, "Modeling emotion in complex stories: the stanford emotional narratives dataset," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 579–594, 2021.
[40] D. Ciuk, A. Troy, and M. Jones, "Measuring emotion: Self-reports vs. physiological indicators," *Physiological Indicators (April 16, 2015)*, 2015.
[41] K. Heimann, M. Nouwens, S. Saggurthi, and P. Dalsgaard, "Micro-phenomenology as a method for studying user experience in human-computer interaction," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–17.
[42] X. L. Cang and et al, "Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition," in *Proc of the 2015 ACM on Int'l Conf on Multimodal Interaction*, 2015, pp. 147–154.
[43] M. M. Jung *et al.*, "Touching the void–introducing cost: corpus of social touch," in *Proc of the 16th Int'l Conf on Multimodal Interaction*, 2014, pp. 120–127.
[44] A. Flagg and K. MacLean, "Affective touch gesture recognition for a furry zoomorphic machine," in *Proc of the 7th Int'l Conf on Tangible, Embedded and Embodied Interaction*, 2013, pp. 25–32.
[45] K. Altun and K. E. MacLean, "Recognizing affect in human touch of a robot," *Pattern Recognition Letters*, vol. 66, pp. 31–40, 2015.
[46] M. A. Conway and C. W. Pleydell-Pearce, "The construction of autobiographical memories in the self-memory system." *Psychological review*, vol. 107, no. 2, p. 261, 2000.
[47] X. L. Cang, R. R. Guerra, P. Bucci, B. Guta, K. MacLean, L. Rodgers, H. Mah, S. Hsu, Q. Feng, C. Zhang *et al.*, "Choose or fuse: Enriching data views with multi-label emotion dynamics," in *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2022, pp. 1–8.
[48] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[49] D. Byrne, "A worked example of braun and clarke's approach to reflexive thematic analysis," *Quality & quantity*, vol. 56, no. 3, pp. 1391–1412, 2022.
[50] P. M. Ullrich and S. K. Lutgendorf, "Journaling about stressful events: Effects of cognitive processing and emotional expression," *Annals of Behavioral Medicine*, vol. 24, no. 3, pp. 244–250, 2002.
[51] R. Hiemstra *et al.*, "Uses and benefits of journal writing," *New directions for adult and continuing education*, vol. 2001, no. 90, p. 19, 2001.
[52] W. Miller, "Interactive journaling as a clinical tool," *Journal of Mental Health Counseling*, vol. 36, no. 1, pp. 31–42, 2014.
[53] E. B. Andrade and J. B. Cohen, "On the consumption of negative feelings," *Journal of Consumer Research*, vol. 34, no. 3, pp. 283–300, 2007.
[54] S. Huron and W. Willett, "Visualizations as data input?" 2021.
[55] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *J Personality & Social Psychology*, vol. 54, no. 6, p. 1063, 1988.
[56] C. Harmon-Jones, B. Bastian, and E. Harmon-Jones, "The discrete emotions questionnaire: A new tool for measuring state self-reported emotions," *PloS one*, vol. 11, no. 8, p. e0159915, 2016.
[57] G. B. Chapman, "Sooner or later: The psychology of intertemporal choice psychology of learning and motivation," 1998, pp. 83–113.
[58] X. Nan and Y. Qin, "How thinking about the future affects our decisions in the present: Effects of time orientation and episodic future thinking on responses to health warning messages," *Human Communication Research*, vol. 45, no. 2, p. 148–168, 2019.

Rubia R. Guerra is a Ph.D. fellow in Computer Science specializing in Data Science and Affective Haptics. They hold an M.Sc. in Computer Science and a B.Sc. in Systems Engineering. Their research focuses on computationally modeling human emotions through touch, with an emphasis on dynamic emotion trajectories and affective touch behaviors in emotionally charged environments. Their work looks into improving the reliability and reproducibility of emotion-based research, with applications in affective computing and human-computer interaction.

Xi Laura Cang is a computer scientist and educator, studying the application of machine learning and deep learning in affective touch and emotionally interactive technologies.

Po-Yu Chen is a Master's student in Applied Computing at the University of Toronto, after graduating with distinction in Engineering Physics from UBC. At the time of this project, Po-Yu was completing a degree in Engineering Physics at UBC. His research focuses on AI in graphic computing, rendering, and physical simulation, with projects on emotion recognition, robotic surgery, and single-cell RNA sequencing. Po-Yu interned at Sony Pictures, where he explored AI to optimize graphic simulations and automate animation. He aims to drive innovation in AI research within computer graphics or computer vision.

Nao Rojas is a M.Sc. student in Computing Science at the University of Alberta, specializing in computer vision and deep learning. At the time of this project, Nao was completing a double degree in Computer Science and Mathematics at UBC. With professional experience spanning computer vision, geospatial data analysis, full-stack development, and product design, they are passionate about finding innovative and user-oriented solutions to environmental and socioeconomic issues. Nao completed their undergraduate studies in Mathematics and Computer Science at the University of British Columbia.

Karon E. MacLean is a computer scientist and mechanical engineer whose research involves haptic technology and affective haptics in human–computer interaction and physical human–robot interaction. She has been at the University of British Columbia since 2000, as a professor of computer science and Canada Research Chair in Interactive Human Systems Design.

SUPPLEMENTARY MATERIALS

S.1 The 'Living Lab' Space

Figure 6 illustrates our data collection space.

Figure 6: Data collection space. The study room was arranged to resemble a comfortable, informal living room to contrast with a sterile laboratory or classroom setting. It was furnished with a couch and other simple decor.

S.2 Data Collection Interface

We developed our own custom user interface (UI) using React and Typescript. Figure 7 illustrate key portions of our UI.

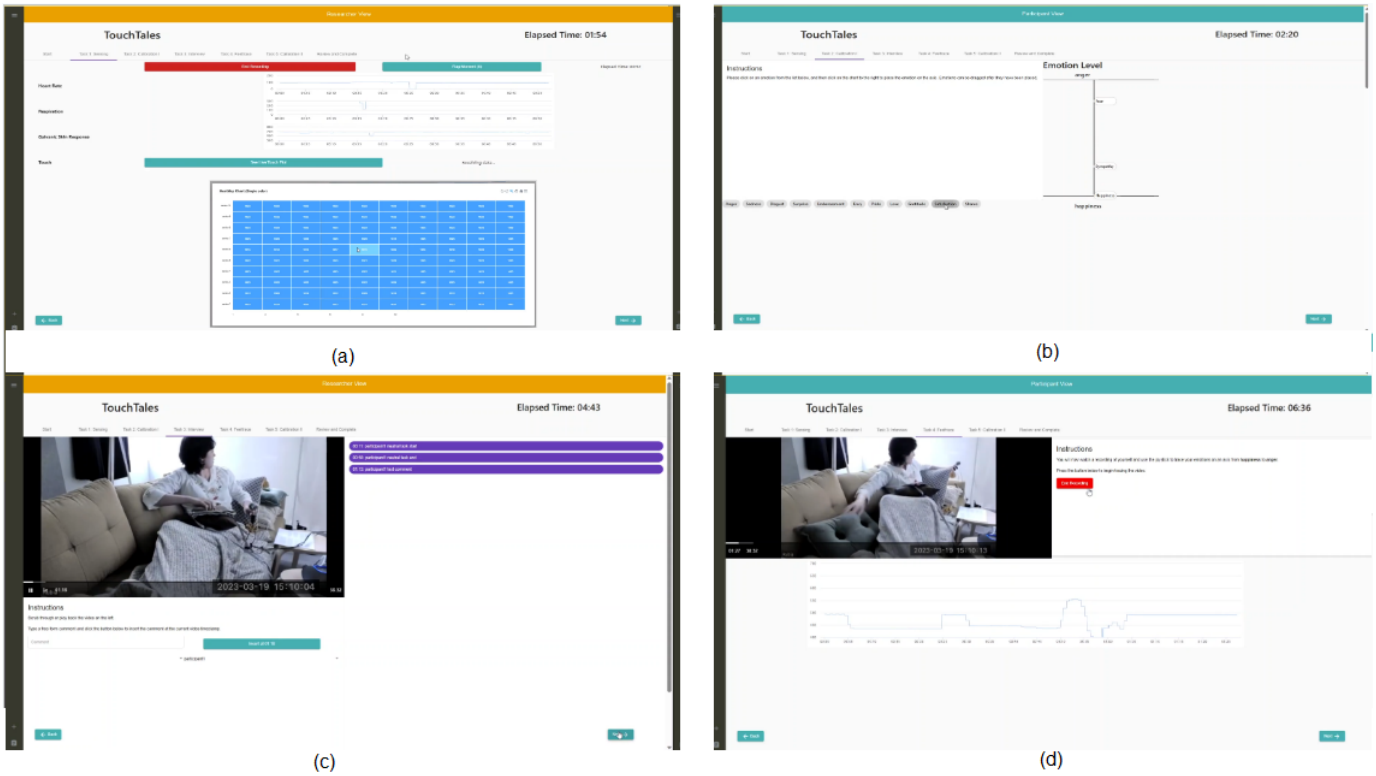


Figure 7: User interfaced designed for data collection, (a) Storytelling recording, with live plotting of physiological and touch data streams, (b) calibration task, (c) interview annotation screen, (d) continuous annotation screen.

S.3 Feeltrace Binning Plots

We processed participant data using our slope-based binning method; the resulting bins are illustrated in Figure 8.

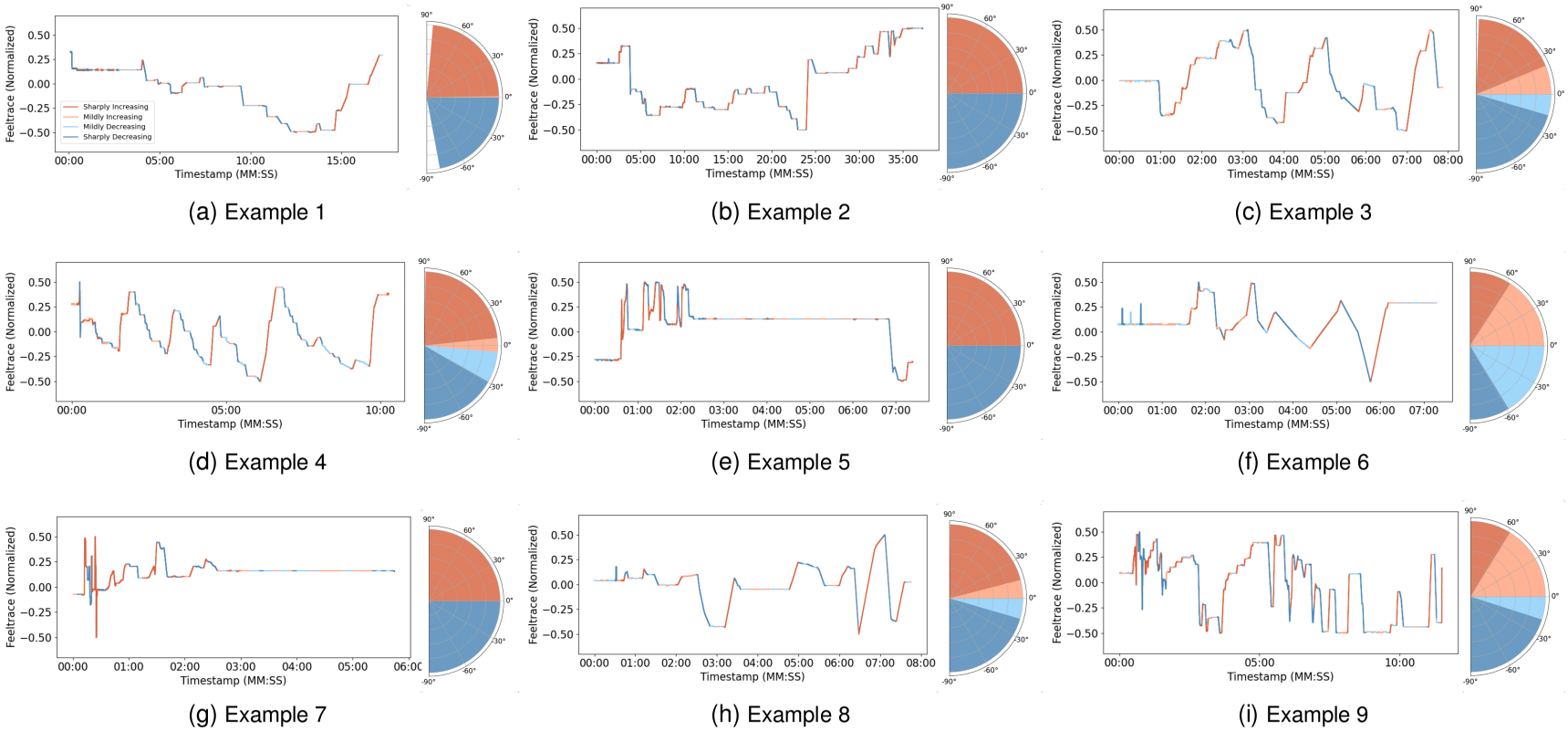


Figure 8: Feeltrace position trajectory (right) and *EmoDir*$_l$ (slope) bin distributions (left). Each panel combines the *position* time series colored by assigned bins for 'example' data (participant number not disclosed for anonymity), and the corresponding polar distribution of the feeltrace.

S.4 Statistical Analysis Results on Machine Learning Experiments

In Section V, we present statistical results of our modelling experiments. Table VII summarizes our analysis results.

Model Information	
Model	MixedLM
Optimizer	Powell
Dependent Variable	Accuracy
No. Observations	810
Method	REML
No. Groups	4
Scale	0.039
Min. Group Size	90
Log-Likelihood	230.69
Max. Group Size	270
Converged	Yes
Mean Group Size	202.50

Table VI: Model information for the updated Generalized Mixed Linear Model (GMLM).

Regression Results					
Variable	**Coef.**	**Std. Err.**	**z**	$P > \|z\|$	**[0.025, 0.975]**
Intercept	0.648	0.029	22.706	**0.000**	[0.592, 0.704]
Session[T.2]	-0.022	0.045	-0.478	0.632	[-0.110, 0.067]
Session[T.3]	-0.005	0.033	-0.159	0.873	[-0.070, 0.059]
Scheme[T.2]	0.019	0.040	0.469	0.639	[-0.060, 0.098]
Scheme[T.3]	0.022	0.040	0.540	0.589	[-0.057, 0.101]
Window[T.5000]	-0.359	0.040	-8.884	**0.000**	[-0.438, -0.280]
Session[T.2]:Scheme[T.2]	-0.005	0.064	-0.073	0.942	[-0.130, 0.120]
Session[T.3]:Scheme[T.2]	-0.011	0.047	-0.243	0.808	[-0.103, 0.080]
Session[T.2]:Scheme[T.3]	-0.003	0.064	-0.050	0.960	[-0.128, 0.122]
Session[T.3]:Scheme[T.3]	-0.018	0.047	-0.381	0.703	[-0.109, 0.074]
Session[T.2]:Window[T.5000]	0.149	0.064	2.331	**0.020**	[0.024, 0.274]
Session[T.3]:Window[T.5000]	0.180	0.047	3.854	**0.000**	[0.088, 0.271]
Scheme[T.2]:Window[T.5000]	-0.036	0.057	-0.624	0.533	[-0.148, 0.076]
Scheme[T.3]:Window[T.5000]	-0.013	0.057	-0.220	0.826	[-0.124, 0.099]
Session[T.2]:Scheme[T.2]:Window[T.5000]	0.007	0.090	0.083	0.934	[-0.169, 0.184]
Session[T.3]:Scheme[T.2]:Window[T.5000]	0.041	0.066	0.617	0.537	[-0.089, 0.170]
Session[T.2]:Scheme[T.3]:Window[T.5000]	-0.001	0.090	-0.014	0.989	[-0.178, 0.176]
Session[T.3]:Scheme[T.3]:Window[T.5000]	0.012	0.066	0.180	0.857	[-0.117, 0.141]
Group Var	0.000				

Table VII: Regression coefficients, standard errors, z-values, p-values, and confidence intervals for the Generalized Mixed Linear Model (MixedLM), with fixed effects for session, scheme, window size, and their interactions (up to three-way). Variables are dummy-coded with the following reference levels: Session = 1, Scheme = 1, and Window = 2000 ms. Only non-baseline levels and their interactions are shown. For example, Session[T.2] represents the contrast between session 2 and the baseline (session 1). Interaction terms (e.g., Session[T.3]:Window[T.5000]) indicate combined effects of deviations from baseline levels.