Prototyping as Epistemic Work: A Collaborative Autoethnography Research Cycle of Prototyping Naturalistic Hand Interaction for Extended Reality

Prashant Rawat[a,*], Sussan T Anukem[a], Rita Orji[a], Ga Wu[a], Francisco Raul Ortega[b], Lizbeth Escobedo Bravo[a], Mayra Donaji Barrera Machuca[c]

[a] *Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada*
[b] *Computer Science, Colorado State University, Ft. Collins, Colorado, United States*
[c] *Department of Computer Science, University of Calgary, Calgary, Alberta, Canada*

Abstract

Human-Computer Interaction designers use prototypes to gain insight into their ideas, determine whether they can be built, and learn about issues with the chosen technology. In this paper, two designers use collaborative autoethnography to document the challenges of building an extended-reality prototype that visualizes neonatal birth-weight data through naturalistc hand interactions. Naturalistic hand interactions enable continuous interaction but are ambiguous and difficult to detect, requiring machine learning (ML) from the prototyping stage. Using a first-person corpus comprising reflective diaries, design artifacts, and 32 meetings recorded over nine months, and analyzed using reflexive thematic analysis, we identified that bringing ML into a prototype changes the design process, as it affects what designers learn from the act of making the prototype. It changes the role of the designer, moving from pure experience design to also a model trainer. These changes stalled the exploration of novel user interactions and affected the final product. Using our experience as a base, we propose guidelines for staging manual, autonomous, and hybrid prototyping and position our findings as first-person observations that motivate future research.

Keywords: Collaborative Autoethnography, Prototyping, Machine Learning, Naturalistic Hand Interaction, First-Person Methods, Data Visualization

*Corresponding author.
Email address: prashant.rawat@dal.ca (Prashant Rawat)

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

1. Introduction

In Human-Computer Interaction (HCI), prototyping is the process of creating an early version of a system, application or interface to test and evaluate its design, functionality and usability. Prototypes are built to learn the advantages and disadvantages of specific technologies, as designers can discover interaction affordances and develop a felt sense of how such technologies might be experienced in use [1, 2, 3]. From this perspective, a prototype is not only a draft of a product but also an instrument of inquiry, of thinking through making and of learning by meeting the resistance of one's material [4]. However, what designers can learn depends on how the prototype is assembled, as different materials and technologies answer different questions while quietly foreclosing others.

There are many types of prototypes, ranging from low-fidelity sketches to high-fidelity designs built with custom software that show how the system will look and behave [1]. For extended reality (XR) technologies, prototypes utilize the same development tools as the final product, e.g., game engines [5, 6] and SDKs [7]. These tools allow designers to create prototypes for specific headsets by easily accessing the system's features, such as hand tracking, controller input, and gaze interaction. Some of these features use pre-trained machine learning (ML) models for gesture recognition and hand tracking [8, 7]. As a result, designers prototype XR interactions that depend on probabilistic inference from the very first step rather than treating recognition as a later engineering concern. However, prior research has shown that pre-trained models introduce uncertainty, latency, and probabilistic errors [9, 10]. Thus, using these tools to create prototypes creates a tension between designer and tool, which introduces new categories of issues that are difficult to anticipate and repair. For example, a gesture may stop being recognized after retraining, or a delay may appear only at run time [11, 12].

In this paper, we treat prototyping of naturalistic hand interaction for XR systems not as background to a system contribution, but as a subject of our own experiences through a collaborative autoethnography (CAE) [13, 14]. Naturalistic hand interactions are continuous, ambiguous, and person-dependent, making them difficult to prototype for. For example, a rule-based recognizer cannot interpret them, and manual simulation can hide the feasibility of the interaction [15, 16]. Consequently, designers of naturalistic hand-based

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

interfaces prototype with ML in the loop, whether they intend to or not. Our first research question (RQ1) is **what can ML prototypes reveal about the feasibility, timing, and robustness of interpreting hand interactions?** Our second research question (RQ2) is **how do these insights change when using manual vs. ML-based hand-interaction interpretation?** Our third research question (RQ3) is **what does collaborative autoethnography make visible about prototyping with ML that other evaluation methods would miss?** To answer these questions, we (the first and second authors) documented our experience of designing and building a naturalistic, hand-based system for visualizing neonatal birthweight data over nine months. The goal of the prototype was practical and implementation-focused: to test whether the hand actions we had designed could be reliably interpreted by the system and used to drive the visualization in real time.

Interpretation is the step that turns a hand movement into a meaningful action; what a movement means is not fixed in advance but settled in context [17]. While building our prototype, interpretation took three forms. They differ not in fidelity but in who performs the interpretive work. At first, we supplied it by hand. Before any model existed, one of us simulated the recognizer: we watched the hand and triggered the response the system was meant to give, following the interaction rules we had set. This is Wizard-of-Oz (WoZ) prototyping in its standard sense, where a person stands in for the envisioned system so an interaction can be tested before it is built [18, 19, 20]. We keep the term WoZ for the technique itself. But in the rest of the paper, we call our use of it **manual scaffolding**, for one reason: WoZ names a study setup in which a user believes the system is autonomous, and we had no users. What we did was WoZ-style triggering as an internal building aid [21]. The term keeps that distinction visible and keeps attention on who performs the interpretive work. As recognizers matured, we delegated interpretation to trained models. We call this **autonomous recognition**. In practice, we settled into a **hybrid**, in which manual scaffolding was selectively reintroduced around the model to keep interaction legible. We do not treat these three as steps toward higher fidelity. They are different arrangements of interpretive authority, and each makes different things learnable for the designer while concealing others.

Our research makes four contributions. First, it offers an account of prototyping with ML in the loop as epistemic work, i.e, as work that changes what the prototype can teach its designers. The account centres on an unexpected

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

observation from our own use, developed in Theme 6: recognition failures extended beyond the interface into how we ourselves read the data, making timing, rather than benchmark accuracy, the constraint that mattered. We scope this as a designer-level observation and identify the user study it motivates. Second, it articulates, from lived design experience, the conditions that produce this effect: the movement between manual and autonomous interpretation, gesture-and-dataset decisions as interaction design, and the migration of design labour into pipelines and iteration loops. Third, it offers practical guidance for staging hybrid, human-scaffolded prototyping that keeps interpretation legible while confronting ML's temporal and probabilistic limits. Fourth, we reflect on CAE as a method for studying partially autonomous systems from within design practice. Such systems often exhibit breakdowns that are intermittent, contextual, and difficult to reproduce through conventional evaluation.

2. Related Work

Our work sits at the intersection of three research threads: reflexive, first-person methods in HCI; naturalistic hand actions as embodied, cognitively grounded interaction; and prototyping practices that span human and ML-driven interpretation. We read all three through a single lens as prototyping as an epistemic activity. We ask how methodological, interactional, and technical choices shape what designers can observe, reason about, and treat as evidence about the interaction they are designing, that is, about its feasibility, its timing, and its behaviour under error [4, 2].

2.1. Reflexive and Autoethnographic Methods in HCI

Ethnography has long shown how technologies are embedded in social, cultural, and organizational life [22, 23], but its reliance on prolonged access and comparatively stable practices sits awkwardly with fast-moving, unstable design work [24, 25]. In response, HCI has increasingly embraced reflexive, first-person methods that treat the researcher's own situated experience as a legitimate and analyzable source of knowledge [14]. Autoethnography and its variants position the researcher as both subject and analyst, connecting lived experience to broader cultural and technical conditions [26, 27, 28]. Within this family, autobiographical design foregrounds sustained personal use of a system to surface experiential insight [29], while CAE brings several researcher-participants together to document, compare, and analyze a shared experience,

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

so that both agreement and disagreement become analytic resources rather than noise to be averaged away [13]. Duoethnography is a different proposition again: it is dialogic and biographical, using the participants' divergent life histories as the site of inquiry to unsettle taken-for-granted meanings [30]. Related through-design and reflective-practice traditions [31, 32, 33, 4] likewise treat the designer's own making and judgment as knowledge in their own right. In HCI, such first-person accounts have proven especially valuable for studying emerging or unstable technologies, from location-based systems to embodied learning and technology non-use [34, 35, 26]. They suit settings where conventional third-person user studies are premature or impractical, and where the phenomenon of interest is design work itself as it unfolds. We position our study within this family. In Section 3, we say why CAE fits an inquiry into shared, role-differentiated design labour, and how it differs from the neighbouring options we did not adopt.

2.2. Naturalistic Hand Actions as an Interaction Challenge

Gesture is integral to cognition. People reason, learn, and make sense of the world in part through embodied action. The hands do genuine cognitive work: they externalize ideas, scaffold problem-solving, and coordinate spatial understanding [36, 37, 38, 39]. A long tradition distinguishes deictic (pointing), iconic (depicting), and metaphoric gestures, each contributing differently to how people communicate and think through spatial and abstract concepts [40, 41, 42]. In HCI, naturalistic hand actions are movements that resemble everyday physical interaction rather than learned command vocabularies. They have been widely explored for 3D manipulation, immersive environments, and data exploration [43, 44]. They are valued for their intuitive mapping to spatial tasks and for lowering the overhead of learning symbolic commands [45, 46, 47, 48, 49]. Yet, the same qualities that make them expressive make them hard to design for. Because they are continuous, ambiguous, and context-dependent, they resist rigid specification and consistent interpretation [50, 51]; distinguishing meaningful intent from incidental movement, maintaining low latency to preserve coupling, and generalizing recognition across users all remain open problems [52, 53]. As situated-action accounts remind us, the meaning of such movement is not fixed in advance but is constituted moment to moment in context [17, 54]. This makes naturalistic hand interaction a demanding case for prototyping. It exposes, in a sharp form, the tension between expressiveness and reliability, and between human interpretive flexibility and computational constraint.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

2.3. Prototyping Across Manual and ML-Driven Interpretation

Prototyping in HCI spans a wide range of fidelity, from sketches, storyboards, and enactment to rule-based mock-ups and fully functional systems [55, 2, 21, 56, 57, 58]. These techniques differ not only in material form but in the questions they make answerable, distributing control, interpretive authority, and uncertainty differently among designer, user, and system [2, 3]. One long-standing move is to supply interpretation by hand: a designer manually simulates behaviour the system cannot yet produce, standing in for it during early exploration [18, 19, 20, 59]. Such manual scaffolding affords responsiveness and interpretive flexibility, and it lets designers probe expressive, ambiguous interaction before committing to a mechanism. But it can also mislead, because a human quietly absorbs the ambiguity that a deployed system would have to handle on its own, leading designers to overestimate feasibility and robustness [15, 16, 60].

Delegating interpretation to a trained recognizer changes the terms of the work. Real-time detection of hand actions becomes possible through spatiotemporal features and deep models [61, 62, 63]. But designers are now exposed to probabilistic error, latency, and dependence on training data. They also face breakdowns, such as misclassification, delay, loss of temporal alignment, that cannot be smoothed over through improvisation because the human is no longer in the interpretive loop [9, 10, 64, 60]. Past research has established that this delegation shifts interpretive authority from person to model, reshapes designer roles and iteration cycles, and makes human-AI interaction distinctively difficult to design well [9, 10, 11, 12]. This literature has largely characterized ML as a difficult design material: unpredictable, opaque, and demanding of new tools and abstractions.

From this line of work, two research gaps emerge, which motivate this paper. First, manual and autonomous interpretation have largely been treated as alternative implementation strategies or as stepping stones on a path toward a finished system, rather than as distinct epistemic arrangements that make different things learnable during design. Second, past work documents ML's uncertainty and timing costs. It says comparatively little about what those costs entail when the interaction itself is an act of sensemaking, where a mistimed or mistaken interpretation can change not just an action but a conclusion. Sensemaking research describes analytic reasoning as a fragile, iterative coupling between a person and their data representation, in which trust and interpretive momentum are easily disrupted [65, 66]. Bringing that sensitivity together with the properties of ML recognition is precisely where

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

our contribution lies, and we approach it from the designer's own vantage point.

3. Collaborative Autoethnography as Method

3.1. Positionality and Reflexivity

Autoethnography asks its authors to be explicit about who they are, because the account is produced through their situated experience rather than despite it [14, 13, 67]. The two of us (first and second authors) are the primary autoethnographers, and we came to the project with different backgrounds. Autoethnographer 1 (A1) holds a bachelor's degree in computer science engineering and a master's degree in data analytics from a UK university, taught computer science at the undergraduate level for five years, and is now completing doctoral work in Canada on multimodal interaction and spatial cognition in XR. Autoethnographer 2 (A2) holds bachelor's and master's degree in HCI, and is pursuing their doctoral degree in visual analytics. A1 works on machine learning, gesture recognition, and spatial thinking and led the recognition pipeline: hand tracking with MediaPipe, a distance-based feature representation, and sequence models for classification. A2 works on interaction and visualization design and led the dashboard and interaction loop, built in Streamlit [1]. Our reflexes differed accordingly. When an interaction broke down, A1 focused on the ML model: more data, a different architecture, or a different threshold. A2 focused on the interaction: a simpler gesture, a clearer mapping, a smaller task. These reflexes are part of our instrument, not noise in it, and this difference between us enriches the study. The same breakdown often reads to A1 as a "model failure" and to A2 as an "interaction breakdown", and the gap between those readings is among our most useful evidence. We state this openly so the reader can weigh our interpretations, and we report our disagreements rather than resolving them away.

3.2. Why Collaborative Autoethnography

Following Chang et al. [13], we adopt an analytical-interpretive CAE approach, as shown in Figure 1. In this approach, collaborative first-person material is systematically analyzed and interpreted using established qualitative methods. In the same manner, Geist-Martin et al. [68] related first-person

[1] https://docs.streamlit.io/

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

accounts to an existing conceptual frame rather than presenting them as a standalone narrative. We chose CAE deliberately over several neighbouring first-person methods, and the reasons are instructive for what the study is and is not.

Our inquiry centres on a shared design phenomenon: the eight-month build itself. We analyze what happened as we built, not our biographies, and this is what separates our CAE from the dialogic juxtaposition of life stories in duoethnography [30]. We do not claim a research-through-design (RtD) contribution. RtD generates knowledge through the deliberate making of an artifact as the research act [31, 32]. Our prototype was built alongside, and partly ahead of, the analytic framing. What we offer is an interpretive reading of that process as a lived design practice. Finally, unlike autobiographical design [29], which centres on a researcher's long-term use of a system, our object is the making of the prototype. What CAE gives us is a way to make shared, role-differentiated design labour analyzable, while keeping every first-person claim tied to an auditable trail.

The study is organized in two stages that preserve traceability from concrete design work to analytic claims. Stage 1 produced a first-person corpus during the build (January - September 2024) by documenting incidents as they occurred, e.g., gesture redefinition, latency thresholds, class pruning, and integration failures. Stage 2 analyzed that corpus using Reflexive Thematic Analysis (RTA) [69]. Throughout, we use breakdown in a specific sense. We draw on the phenomenological tradition in which a tool becomes visible precisely when it stops working smoothly [70, 4]. A breakdown is a moment when the prototype ceased to behave as we expected and thereby became available for reflection: a misrecognition, a delay, a gesture that could not be operationalized, a mapping that failed in real time.

We also state plainly what we mean by *epistemic work*, and what we tracked. We use the term in the established sense of prototyping as inquiry: activity that produces knowledge for the designers [4, 2, 1]. Prototyping with ML in the loop did not only change our design. It changed (1) the design decisions we could make, (2) our research questions and our definition of the problem, (3) our roles and the division of labor between us, (4) what we counted as a successful outcome, and (5) what the prototype could be treated as evidence for. Table 2 tracks this learning across the build: for each phase it names the CAE record we kept, an example reflection, what we learned (the phase's output), and the design change that became the next phase's input. Breakdowns matter analytically because they are exactly where the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

prototype revealed what it could, and could not, be treated as evidence for.

3.3. Study Context, Timeline, and Ethics

The work ran for nine months. A1 led the dataset and the recognizer and was the primary autoethnographer. This work began in January 2024, and real-time recognition came together in April. A2 started building the dashboard in February and integrated recognition into the web prototype from May. Our backgrounds are described in Section 3.1. Throughout, we consulted domain experts and logged their input: an expert in gesture and 3D user interface design, and an expert in machine learning. We met weekly with one advisor on implementation and monthly with the wider group on specific problems. Following guidance on ethics in CAE [71], we report our own experience and protect the identities of the external expert consultants. No personal data appears in our corpus. The materials behind this account, the development log, reflection prompts, trained model, and inference notebook, are provided as supplementary material so the research cycle can be inspected.

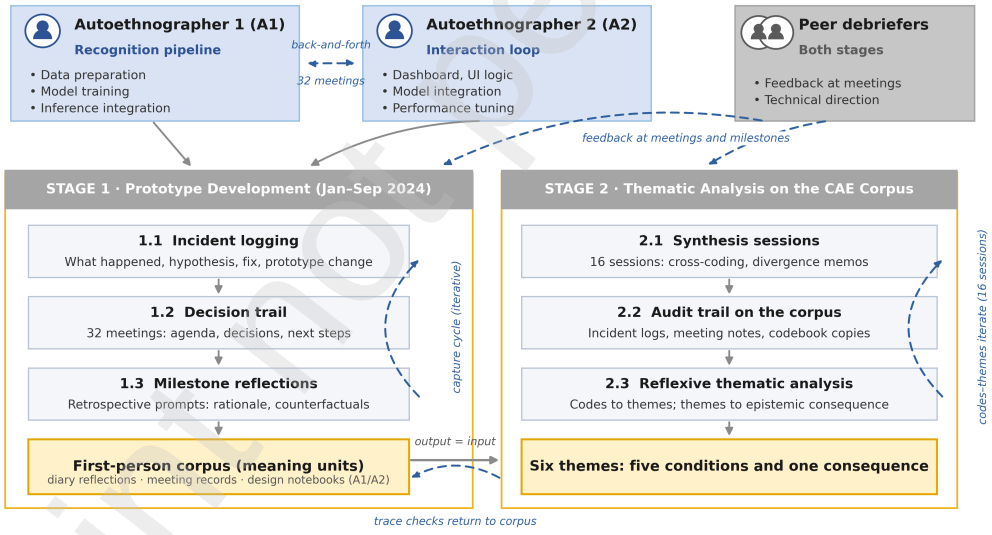


Figure 1: Our analytical-interpretive collaborative autoethnography. Stage 1, the build (Jan–Sep 2024): our two roles feed a fixed capture protocol that produces the first-person corpus. Stage 2, reflexive thematic analysis: independent coding, cross-coding with divergence memos, and synthesis sessions produce six themes. The curved dashed arrows mark the iterative loops: building and capturing ran in a cycle, and coding and synthesis iterated until themes stabilized.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

3.4. Team Roles and Data Ownership

Our roles are described in Section 3.1; here we note what matters for the corpus. Because our responsibilities differed, we encountered different breakdowns first-hand and observed others second-hand. This asymmetry is a designed feature of the study: it lets us distinguish direct from reported experience and keeps the role-specific readings described above traceable to their owner [13, 72]. The other authors participated as peer debriefers, offering feedback on framing and technical direction, but did not contribute autoethnographic material to the analytic corpus.

3.5. The CAE Corpus

Table 1 summarizes the corpus analyzed in Stage 2. All counts refer to artifacts produced during the build.

Table 1: CAE corpus summary and counts.

Artifact	**Count**	**Purpose in the corpus**
Structured diary reflections (A1)	continuous	Long-form reflections anchored to specific incidents, constraints, and decisions.
Structured diary reflections (A2)	continuous	Parallel long-form reflections; used to surface perspective differences.
Development + CAE meetings	32	Notes and action items; avg. meeting time 30 min; used as the primary decision trail.
Cross-coded meaning units	16	Shared subset coded by both A1 and A2 to calibrate interpretation and expose disagreement.
Synthesis sessions (analysis meetings)	16	Code comparison, memo writing, theme refinement, and trace checks (Stage 2).
Design notebooks (A1/A2)	see Figure 2	Contemporaneous notes, sketches, incident logs, and implementation rationale.
Retrospective prompts (A1/A2)	see Table 3	Structured recall of goals, breakdowns, and alternatives at key milestones.

3.6. Data Capture Protocol During Development

We captured material consistently as we worked, both to reduce hindsight bias and to support later traceability, in keeping with CAE's emphasis on systematic documentation of lived experience [13]. Three practices structured the record: (1) incident logging, (2) a meeting decision trail, and (3) milestone reflections. Through incident logging, whenever a breakdown or high-leverage decision occurred, such as a gesture-definition change, a latency threshold, or a

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

class pruning, the responsible author recorded four fields in the incidents logs: what happened, an immediate hypothesis about why, the fix we attempted, and what changed in the prototype as a result. Using a meeting decision trail, each of our 32 development-and-analysis meetings was recorded, including its agenda, decisions, and next steps. When a decision altered the interaction loop (the chain from hand movement, to recognized action, to visualization response), such as confidence thresholding, smoothing, or a fallback behaviour, we recorded the intended behavioural change in the decision trail (for example, "raise the confidence threshold so false triggers stop"), so that later effects could be traced to their cause. Finally, through milestone reflections, at key junctures we wrote longer diary and retrospective entries to capture deeper rationale, emotional load, and the counterfactuals we had weighed but not taken, guided by the structured prompts in Table 3. For example, what might have happened had we kept the interaction in a manually scaffolded mode longer. Together these practices kept the corpus close to the lived texture of the work rather than a sanitized reconstruction of it.

3.7. Unit of Analysis

For Stage 2, we segmented the corpus into meaning units, e.g, bounded segments of the record, each centred on a single breakdown, decision, or meeting segment corresponding to one decision. Units came from every phase of the build (Phases 1-4), so the analysis covers dataset work, interaction design, implementation, and integration alike. We retained a unit when it either (a) changed the interaction, the gesture vocabulary, or the feedback timing, or (b) changed what the prototype was treated as evidence for in subsequent design moves. This second criterion is deliberately epistemic as it selects for the moments in which not just the artifact but our understanding of what the artifact could tell us shifted.

3.8. Audit Trail and Analytic Accountability

We maintained an audit trail to keep every claim traceable to the corpus, aligned with CAE's emphasis on transparency and shared analytic accountability [13, 26], e.g., dated meeting notes (32 entries), incident logs and notebook anchors (examples of these contemporaneous notebooks appear in Figure 2), versioned codebooks from Stage 2, and synthesis memos from 16 sessions. Every theme we report is grounded in multiple meaning units drawn from multiple artifact types and can be traced back through these materials. Because our aim is interpretive depth rather than consensus, we did not treat

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

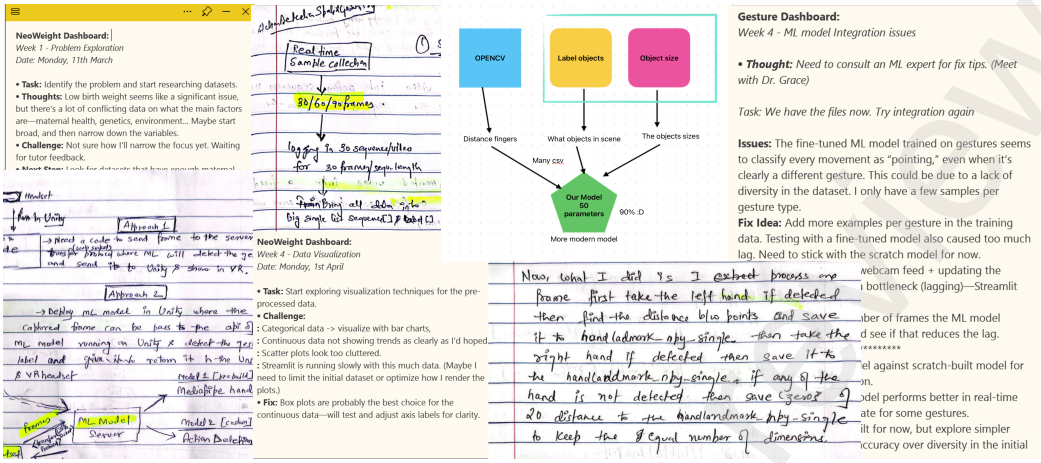


Figure 2: Pictures of the notes taken during the CAE process

coder agreement as a measure of validity; instead, as we describe below, we preserved disagreement as evidence.

4. Stage 1: The Build

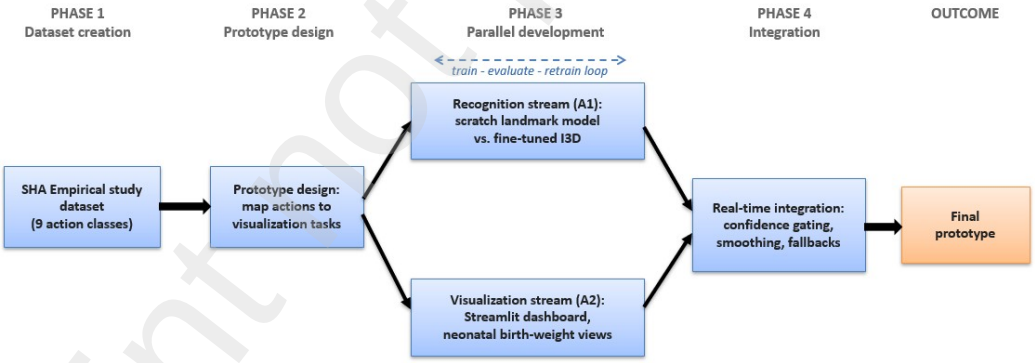


Figure 3: The prototype development process across the four phases of the build. The Spatial Hand Actions dataset in Phase 1 comes from Rawat and Barrera Machuca [73]. Later phases cover interaction design (mapping actions to visualization tasks), the parallel implementation of the recognition pipeline and the web dashboard, and real-time integration.

This section reports Stage 1 of the study: the build itself. We report it as a sequence of phases and mark where the interpretation shifted among

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

Table 2: How learning moved through the build. For each phase: the CAE record we kept, an example reflection from the corpus, what we learned (the phase's output), and the design change that became the input to the next phase.

Stage of development	**CAE record**	**Example reflection**	**What we learned (output)**	**Design change fed forward (input)**
Phase 1: Dataset adoption and action semantics	Incident log; meeting decision trail	A1: "uneasy reducing rich, expressive movement to a simplified taxonomy"	Class boundaries fix the interaction vocabulary; confusable classes cannot both survive	Pruned and merged class set; constraints on which actions the design could use
Phase 2: Interaction design	Meeting decision trail; milestone reflection	A2: "are we rehearsing a fiction, or designing something that could live in practice?"	The action-to-task bindings decide what the prototype can test at all	Fixed action-to-task mapping; requirement for fallback paths
Phase 3: Parallel implementation	Incident log; diaries	A1: "high accuracy on paper collapsed into frustration once lag made the hand untrustworthy"	Temporal alignment, not offline accuracy, governs whether the interaction survives	Deployed landmark model; 70-frame window; gating and smoothing parameters
Phase 4: Integration and use	Milestone reflections; meeting decision trail	A2: "every failed recognition risked pulling us away from what the numbers were telling us"	Recognition error disturbed our own reading of the data	Hybrid fallback paths; the user study this account motivates

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

manual scaffolding, autonomous recognition, and hybrid. It is important to note that these moves were not planned against a fixed methodology. Each was triggered by an emergent need, most often a breakdown that the current arrangement could no longer absorb, and was logged at the point it occurred, so that the arrangement in use is always tied to an observable event rather than asserted after the fact. The build ran in four phases, as shown in Figure 3: dataset adoption and action semantics; interaction design and task mapping; parallel implementation; and real-time integration.

4.1. Phase 1: Dataset Adoption and Action Semantics

We built the study on the Spatial Hand Actions (SHA) dataset [73], which was collected from an empirical study. In this dataset, 18 participants completed four 3D assembly tasks, two block-construction and two Lego, while thinking aloud. Rawat and Barrera Machuca [73] captured the non-gestural hand movements that accompany spatial reasoning and named them Spatial Hand Actions. Concretely, the dataset consists of short video clips of the participants' hands, recorded from an overhead camera above the tabletop while they manipulated the wooden blocks and Lego pieces of the assembly tasks. Each clip is labelled with one of nine action classes, grouped into three categories: spatial relation actions (Comparison, Hand Relation, Pointing), spatial visualization actions (Grabbing, Inspect Object, Placing Object, Separating Object), and spatial orientation actions (Mental Image, Orientation); Figure 7(a) illustrates all nine by category. These are not command gestures. They are the movements people produce naturally while reasoning about objects in space: holding two pieces side by side to compare them, rotating a part to reorient it, tracing an outline in the air while imagining an assembly. Because every action in the dataset was performed on physical pieces, the pieces are part of the vocabulary the recognizer learned. This shapes how the prototype is used, as we explain in Section 4.2.

We adopted this dataset for a specific reason: exploring data is itself a spatial-reasoning task. A person compares magnitudes, tracks outliers, and moves between overview and detail, and the SHA actions externalize exactly this kind of thinking with the hands [74, 75]. A vocabulary of actions that accompanies spatial reasoning during assembly is therefore a plausible vocabulary for driving a visualization whose use demands the same reasoning. We reviewed the raw videos from the dataset, discarded corrupted or out-of-frame clips, standardized the format, and split the clips by class. In the present work, this dataset served two roles at once: training material for the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

recognizer and a working specification of the action vocabulary the interaction could use. Treating the dataset work as interaction design rather than as pre-processing, we logged the moments where action semantics were contested. Some were category-boundary disputes, e.g., whether to keep fine-grained distinctions between similar movements or merge them so a model could learn them, which we recorded as negotiation episodes because they were, in effect, arguments about what the interaction would be allowed to mean. Other moments concerned segmentation: what counted as the start and end of an action when a stream of continuous movement had to be cut into discrete, recognizable units. Still others were vocabulary pressure, as early recognition failures that pushed us to prune or redefine hand actions before we had even integrated them. Even at this stage, before a single gesture drove the interface, we could feel the pull toward discretization: the quiet conversion of a rich, continuous repertoire into a small set of learnable classes, a pressure that prior work on ML as a design material also reports [10, 12]. We began to treat that pull itself as a finding rather than a chore.

4.2. Phase 2: Interaction Design and Task Mapping

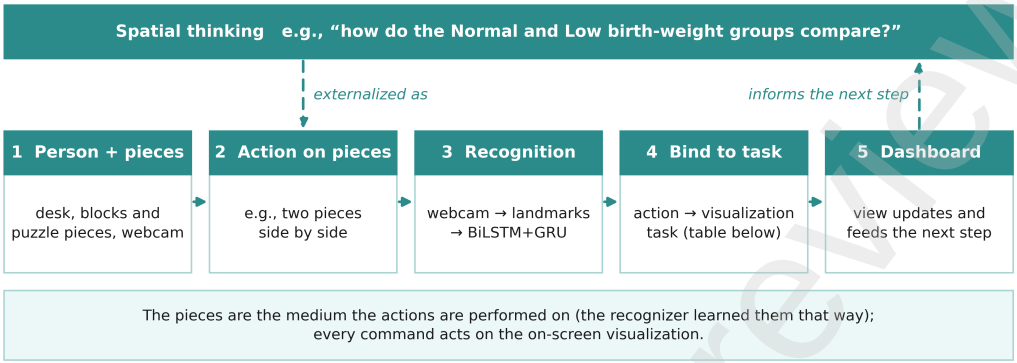
In parallel, we designed a visualization to explore determinants of neonatal birth weight, such as maternal health, age, prenatal development, and socioeconomic factors. We fixed the interaction loop before any automation entered it: which visualization states exist, which analytic tasks are supported (filter, compare, zoom into a distribution's tail, change view), and which hand actions map to those tasks. Sketches, flow notes, and mapping rationales in our notebooks recorded decisions about which interaction would later constrain recognition, which needed continuous control versus a discrete trigger, and which were especially sensitive to latency. Figure 4 shows examples of the hand actions and the prototype they drive. Wherever an interaction demanded immediate responsiveness, repairability under ambiguity, or continuity under partial failure, we flagged it as a candidate for manual scaffolding or a hybrid fallback during integration. In retrospect, these annotations were an early, tacit acknowledgment that not every interaction could safely be handed to a model.

We matched each action's spatial semantics to established taxonomies of visualization tasks from the interaction categories of Yi et al. [76] such as select, filter, and explore, and Brehmer and Munzner [77]'s task typology. For example, Grabbing maps to filtering because both single out a subset to hold onto, and Hand Relation maps to zooming because the distance between two

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

(a) One action, end to end



(b) Action-to-task mapping

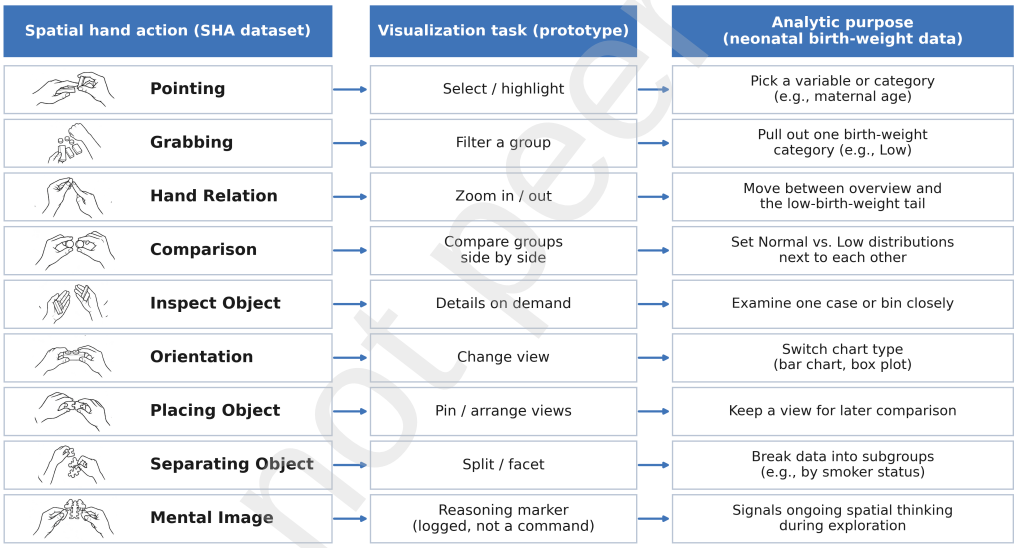


Figure 4: How hand actions drive the visualization. (a) One action end to end: the analyst's spatial reasoning is externalized as a naturalistic action on physical pieces, recognized from the webcam, bound to a visualization task, and answered by the dashboard, whose response feeds the next reasoning step. (b) The full mapping of the nine SHA actions [73] to visualization tasks and their analytic purpose for the neonatal birth-weight data. The puzzle pieces in the illustrations are also used in the prototype.

hands externalizes a change of scale. The Hand Relation-to-zoom binding is visible in the running prototype, and the remaining bindings were fixed at design time in this phase.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

Three points need to be stated plainly here, because the illustrations in Figure 4 and Figure 7 raise them. The first is what is in the person's hands. A person uses the prototype at a desk with a few wooden blocks and puzzle pieces in front of a standard webcam, and handles them the way the participants in the dataset did. The pieces are the physical medium through which the naturalistic actions are performed; they are not data objects. Any piece serves, and the action rather than the piece carries the command. This is a design decision, because the recognizer learned its vocabulary from people handling assembly pieces, and without pieces the same movements would collapse into air gestures, a different vocabulary from the one the model knows.

The second point is what each action refers to. The referent of every command is the on-screen visualization, never the piece. Figure 4(a) follows one action end to end, from the analyst's reasoning, through the pieces and the recognizer, to the dashboard's response, and Figure 4(b) gives the full mapping of the nine actions to visualization tasks. One action, Mental Image, issues no command at all; it is detected and logged as a marker that spatial reasoning is under way.

The third point is why this pairing is natural rather than arbitrary. Each binding pairs a physical operation with the analytic operation it resembles: separating pieces and faceting data both break one thing into parts, holding two pieces together and comparing two distributions both put things side by side, changing the distance between the hands and zooming both change scale. The person does not learn a gesture set. They handle the pieces as they would while reasoning about any spatial problem, and the visualization follows.

4.3. Phase 3: Parallel Implementation

Implementation proceeded in two streams with shared checkpoints in our meetings. On the recognition side, A1 built and compared two pipelines. The first worked from hand landmarks. For each frame, MediaPipe Hands [78] returned 21 keypoints per hand. From these, A1 computed 20 pairwise distances per hand, such as thumb tip to each fingertip, wrist to each fingertip, and joint spacings that show bending. The two hands together gave a 40-dimensional feature per frame. When no hand was detected, its 20 values were set to zero, so the feature retained the same size. A1 first tried min-max scaling, but it made the model too sensitive to small changes, so the distances were normalized instead. A sequence model built from BiLSTM and GRU

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

layers [79, 63] then classified short windows of frames; it was kept shallow to limit overfitting on a small dataset, and trained with dropout and early stopping.

The second pipeline fine-tuned I3D [80], pretrained on CharadesEgo [81]. A1 froze the early layers, trained only the last few, and left out flips and similar augmentations, because the left and right hands play different roles. In real-time tests, the two traded off. The I3D model had faster raw inference, near 50-60 ms per frame against more than 200 ms for the scratch model, but it leaned toward a few labels and lagged in use, so the landmark model was the one wired into the live prototype. None of this was neutral pre-processing. Choosing which distances to measure, how to normalize them, and what to do when a hand disappears already fixed what a gesture was to the system, and so what the model could later be asked to recognize. We tracked training across epochs and inspected per-class behaviour, as shown in Figure 5 and Figure 6.

We include these results not as the paper's contribution but as context. Both pipelines reached usable yet imperfect recognition, in the range that is typical for nine-class naturalistic action recognition trained on a small dataset, and that imperfection is the condition our account studies rather than a shortfall to be explained away. It is precisely their imperfection such as per-class variability, sensitivity to segmentation, inference latency, that made thresholds, smoothing, and our own expectations about timing into design material. On the visualization side, A2 built the web dashboard, the prototype's visualization interface, in Streamlit. It supports categorical comparison, distribution inspection, and filtering, and tracked responsiveness issues (e.g., render latency under filtering) that shaped how the interaction felt long before any formal measure would have flagged them.

4.4. Phase 4: Real-time Integration and Hybrid Scaffolding

After integrating the hand recognition ML model into the dashboard (the Streamlit visualization interface built in Phase 3 and shown in Figure 7(b)), we turned our attention to stabilizing interaction under ML uncertainty. We logged the decisions that altered interaction behaviour. These included confidence gating to suppress spurious triggers and temporal smoothing to stabilize state changes across a window of predictions. They also included fallback behaviours, in which the system ignored low-confidence output, requested confirmation, or reverted to a non-gesture control path to preserve the continuity of exploration. This is what hybrid means concretely in our

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

systems. In other words, the selective reintroduction of manual scaffolding to protect interpretive continuity without discarding the model's behaviour. Integration produced the densest run of meaning units in the whole build, because it was here that timing, misrecognition and trust breakdowns became visible all at once and forced us to decide, repeatedly and explicitly, what the prototype could and could not be treated as evidence for (Figure 7).

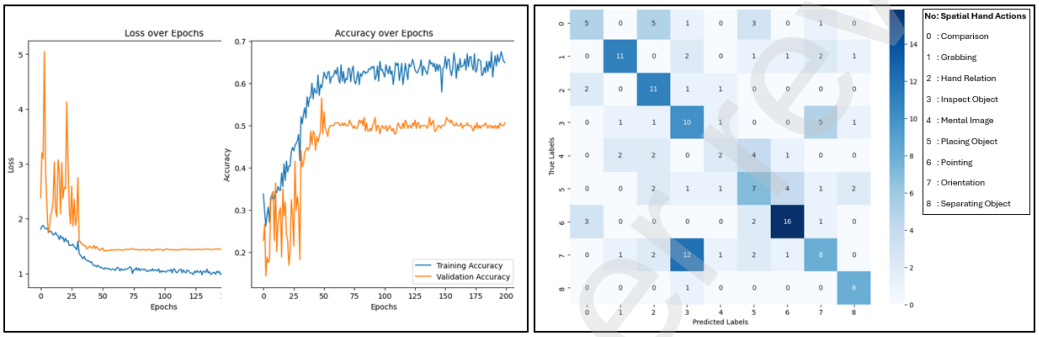


Figure 5: Performance of the hybrid BiLSTM–GRU gesture recognition model over training epochs. Left: loss and accuracy curves for training and validation. Right: confusion matrix across nine SHA [73] classes

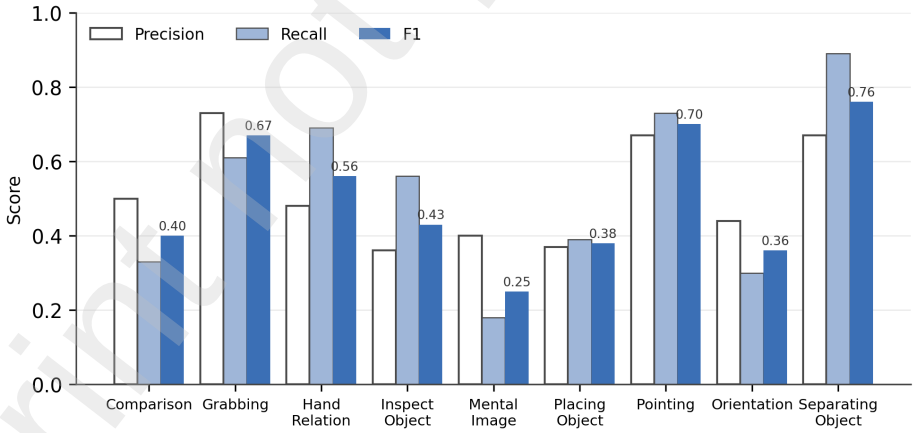


Figure 6: Individual Spatial Hand Action Classification Score Chart for the Fine-Tuned I3D Model on the SHA Dataset

Figure 7 shows this coupling at run time. The camera feed passes through MediaPipe to produce hand landmarks. A sliding 70-frame window of distance

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

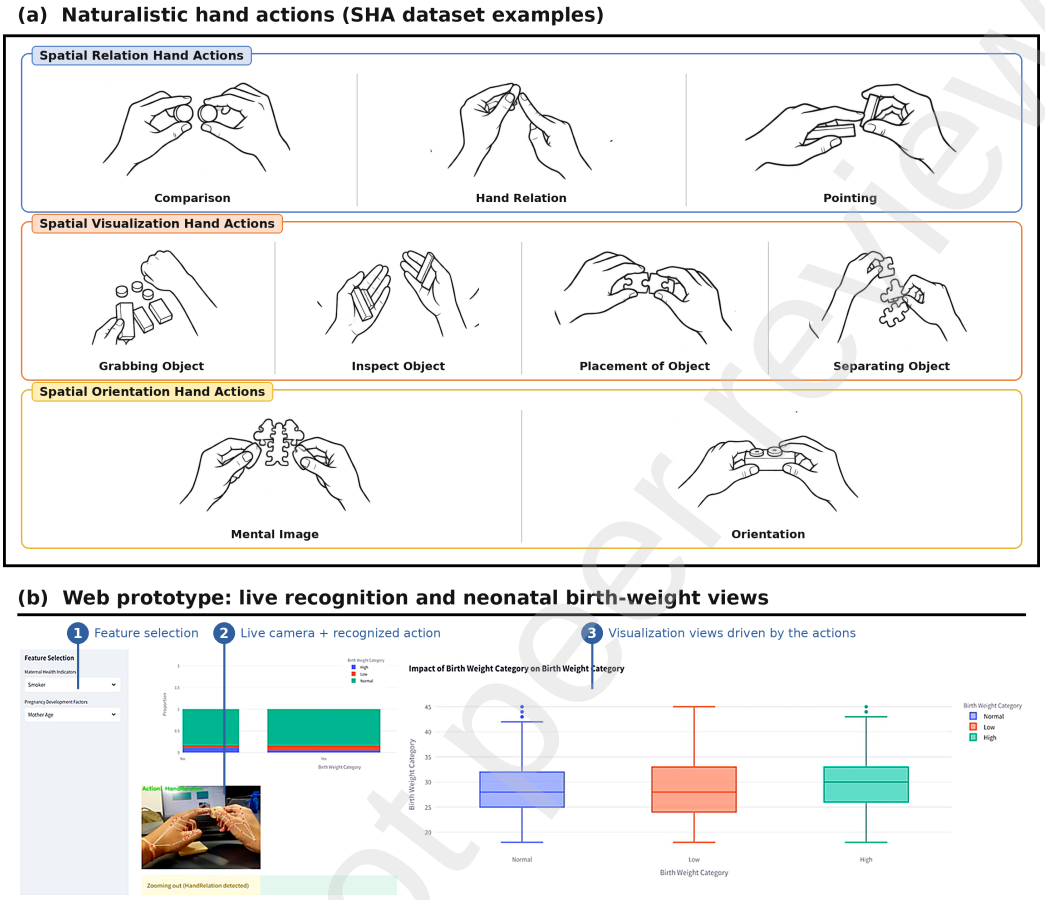


Figure 7: The dataset Example and the prototype. (a) The nine spatial hand actions of the SHA dataset [73], grouped by category: spatial relation (top row), spatial visualization (middle row), and spatial orientation (bottom row), each performed on assembly pieces. (b) The web prototype in use: (1) feature-selection panel for the neonatal birth-weight data; (2) live webcam view, in which the person is handling a wooden piece while the system reports "Zooming out — Hand Relation detected"; (3) the visualization views the actions drive. The same kind of pieces appear in (a) and (b): the person operates the prototype by handling them, and each recognized action issues a command on the on-screen views (Figure 4).

features goes to the recognizer. If it survives confidence gating and smoothing of this recognized action, then it issues exactly one visualization command and the dashboard updates in place. The actions in use mirror what an analyst does with the data e.g., Hand Relation widens or narrows the view

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

the way zooming does, Grabbing pulls out one birth-weight category, and Comparison sets two distributions side by side (Figure 4).

Table 3: CAE reflection prompts used during Stage 1. Prompts were designed to elicit traceable accounts of design decisions, breakdowns, and shifts in prototyping work.

No.	Remembering (situated reconstruction)	Reflection (analytic articulation)
1	What were you attempting to prototype or test at this moment (interaction, gesture, pipeline, or behavior)?	What kind of knowledge did you expect this prototype configuration to give you?
2	What breakdowns, frictions, or unexpected behaviors occurred during this step?	How did these breakdowns reshape what you could or could not learn from the prototype?
3	What design decisions did you make in response (e.g., simplifying gestures, changing thresholds, adding scaffolding)?	What trade-offs did these decisions introduce (e.g., expressivity vs. robustness, control vs. autonomy)?
4	What form of prototyping was active here (manual scaffolding, autonomous recognition, or hybrid)?	How did this arrangement shape your sense of control, trust, or confidence in the interaction evidence?
5	What labor became visible or dominant at this point (e.g., data work, tuning, integration, interface design)?	How did this redistribution of effort affect the pace, direction, or focus of the design process?
6	Looking back, what alternatives were considered but rejected at this stage?	What would you stage differently next time to preserve learning without prematurely constraining the interaction space?

5. Stage 2: Reflexive Thematic Analysis

Stage 1 produced the corpus; this section reports how we analyzed it. We analyzed the corpus (Table 1) using RTA [69], chosen because its commitment to interpretation over consensus aligns with a study whose aim is to make situated design reasoning legible rather than to measure agreement. Our goal was an auditable account of how our prototyping work changed as recognition entered the interaction loop, and of which trade-offs and forms of labour became consequential.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

5.1. Procedure

Analysis proceeded through an explicit, recorded pipeline. We first familiarized ourselves by independently rereading the full corpus and reconstructing a shared timeline anchored to shifts in interpretation (between manual, autonomous, and hybrid arrangements), changes in gesture-vocabulary, and integration incidents such as latency and misclassification. We first independently coded our own material. The shared meeting records, using both incident-led codes (e.g., "confidence gating to suppress spurious triggers", "class pruning after repeated misrecognition", "UI responsiveness constrains gesture mappings"), and reflective ones ("loss of interpretive control","uncertainty becomes interaction material"), each carrying a pointer back to a dated artifact. To calibrate and, more importantly, to surface productive divergence, we cross-coded a shared subset of 16 meaning units spanning all four phases and all three arrangements; the aim was never inter-coder reliability but the visibility of interpretive difference. Across 16 synthesis sessions, we clustered codes into candidate themes, wrote analytic memos relating each theme to the movement of interpretive authority, and ran trace checks confirming that every theme rested on multiple meaning units across more than one artifact type. Finally, we refined and named themes by testing whether each explained incidents across phases, captured a recurring trade-off or redistribution of labour, and carried an epistemic consequence, i.e., changed what the prototype could be treated as valid evidence for.

Handling disagreement as data. When the two of us read the same meaning unit differently, we did not adjudicate one reading as correct. Instead, we wrote a short evidence memo recording both interpretations, the assumption that differed (often learnability versus expressivity, or robustness versus richness), and the design consequence that followed: pruning a gesture class versus introducing a hybrid confirmation step. These memos became some of our sharpest evidence for how role and disciplinary lens shape what a prototype is taken to show. Table 4 presents two representative traces from the meaning unit through code to the theme and its epistemic consequence.

5.2. Themes

RTA produced six themes. Five of them (Themes 1-5) name conditions that hand recognition imposed on our prototyping. The sixth (Theme 6) names what happened when those conditions converged during our own use of the prototype. We present each theme with material from both authors

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

Table 4: Two example traces from meaning units to codes and themes (Stage 2).

Meaning unit	**Initial code(s)**	**Theme**	**Epistemic consequence for prototyping**
A1: "When I switched from WoZ to ML, I felt like I had lost control over the prototype... the unpredictability constantly disrupted the flow."	Loss of interpretive control; autonomy exposes breakdowns; interaction flow becomes contingent	**Where interpretation sits changes what "working" means**	Evidence shifts from "Can this interaction feel fluid?" to "Under what uncertainty/latency does this interaction remain learnable and trustworthy?"
A1: "Most of my time was spent cleaning mislabeled sequences and rethinking what counted as a gesture."	Data/pipeline labor as interaction design; redefining action boundaries; discretization pressure	**Gesture and dataset work is interaction design**	Evidence shifts from "Does this mapping make sense?" to "Which actions can be operationalized without collapsing naturalistic richness?"

and from more than one record type. We develop each theme through our own reflections and read them against the literature as we go.

5.2.1. Theme 1: Where Interpretation Sits Changes What "Working" Means

This theme is grounded in 12 meaning units across diaries (A1: 5, A2: 7), incident logs, and meeting records.

> *May 2024: I (A2) have a gesture that works. When A1 manually triggers the response, the dashboard filters the data cleanly and the interaction feels right. We connect the model, and the same gesture stops working. The model reads it as something else, then reads it correctly, then drops it. I do it again, slower. It still fails. The interaction I just watched succeeded a minute ago is now unreliable, and I cannot fix it by trying harder, because I am no longer the one deciding what the movement means.*

When one of us stood in for the recognizer, ambiguity was a resource: an unclear movement could be read on the spot, and the interaction carried on [18, 19]. Once the model took over, that same ambiguity returned as misclassification and timing errors that no improvisation could absorb, because interpretation now sat with the model. A2 recorded the discontinuity plainly:

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

"gestures that worked flawlessly when we triggered them by hand suddenly failed when the model was in charge". This is the transfer of interpretive authority that past research describes [9, 10], but living it made a further point clear: what changed was not only the interaction but our test for whether it worked. This is why we kept adding hybrids like confidence gating, confirmation steps, smoothing, and fallbacks, that returned interpretation to a person selectively rather than all at once. A1 reached the same view from the model side, proposing that we keep a manual path until the model was mature enough to be trusted on its own. **Implication:** Hybrid scaffolding is not a stage on the way to full autonomy. It is a way to protect the continuity of interpretation while still exposing ML uncertainty as design material.

5.2.2. Theme 2: Gesture and Dataset Work as Interaction Design

This theme is grounded in 15 meaning units across diaries (A1: 7, A2: 8), incident logs, and meeting records.

> *February 2024: A1 and I (A2) are deciding what counts as a gesture. On A1's side, each hand has become a list of 20 distances, normalized, and a model learns to tell a small set of actions apart. Two of the actions keep getting confused. A1 wants to merge them so the model can learn the boundary. I keep thinking about the person who will make the movement, and how the two actions are not the same even if they look close to a model. In effect, we are deciding what the interaction is allowed to mean before anyone has used it. I write in my notes: are we rehearsing a fiction, or designing something that could live in practice?*

The dataset played two roles at once: it trained the recognizer, and it fixed the vocabulary of actions the interaction could use. Decisions we might have filed under data work, such as which distances to measure, where one action ended, and the next began, whether to keep or merge a class, set the boundaries of the interaction itself. The tension was felt on both sides. A1 found that mapping spontaneous, continuous hand movement onto a fixed set of classes introduced inconsistencies that made recognition unreliable. Even so, A1 recorded feeling "uneasy reducing rich, expressive movement to a simplified taxonomy". A2 accepted the other horn of the dilemma: "without categories the model would choke". To cut a continuous, situated repertoire into learnable classes [17] is not preparation for design; it is a design decision that narrows, in advance, what the interaction can express [10]. **Implication**: The design space becomes a function of the dataset structure. Discretization is an epistemic commitment, not only a technical one, and it should be made

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

in the open.

5.2.3. Theme 3: Timing, Not Accuracy, Governs Thinking-Through-Action

This theme is grounded in 9 meaning units across diaries (A1: 6, A2: 3), incident logs, and meeting records.

> *May 2024: The model scores well on our test data, so I (A1) expect the live version to feel good. It does not. To match the length of the sequences the model was trained on, the live loop waits until it has collected 70 frames before making a prediction. On-screen, this means the hand moves and the system answers a beat late every time. A quick action blurs across the window and comes back as the wrong label. I raise and lower the window length, trade a little accuracy for a little speed, and watch the same trade play out: when the response lags, I stop trusting my own hand.*

Offline accuracy did not predict whether an interaction was usable. Small timing gaps as a late response, an intermittent miss, broke the sense that the system was following the hand as it moved, which is the coupling that embodied, thinking-through-action depends on [37, 54]. A1 wrote that "high accuracy on paper collapsed into frustration once lag made the hand untrustworthy". A2 asked the question that reordered our priorities: "why chase accuracy when the interaction logic itself breaks under delay?" We began to treat the window length, smoothing, and gating as interaction parameters in their own right. We tuned for feel rather than for a benchmark. **Implication:** In these ML-based prototypes, latency is not a debt to be paid off later. It decides what kind of interaction evidence the prototype can produce at all.

5.2.4. Theme 4: Labour Migrates into Pipelines and Infrastructure

This theme is grounded in 7 meaning units across diaries (A1: 2, A2: 5), incident logs, and meeting records.

> *February-March 2024: I (A2) meant to spend this stretch on the dashboard. Instead, I spend it on the data: cleaning it, normalizing it, deciding what to do with missing and inconsistent entries. Each time the underlying dataset changes, a chart I had working breaks, and I end up going back to fix the visualization rather than improving it. The design work I set out to do keeps getting pushed behind the work that keeps the data usable. It feels like the interface is waiting on the pipeline.*

As recognition became central, the centre of gravity of the work moved with it. More of the effort went into the conditions under which interaction

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

could be tested at all, such as preparing data, iterating the model, and holding the runtime stable, than into the interaction itself. A2 named the dependency directly: "every time the dataset shifted, my visualization broke; the UI felt hostage to the pipeline". This is the redistribution of design labour that past work anticipates [9, 64], but experiencing it showed its cost to knowledge. When the infrastructure is unstable, the prototype cannot answer questions about the interaction because a breakdown might be due to the interaction or to the plumbing beneath it.

Implication: The prototype's value as evidence came to depend on the stability of the pipeline, not on the interface logic alone.

5.2.5. Theme 5: Iteration Loops Can Stall Exploration

This theme is grounded in 11 meaning units across diaries (A1: 4, A2: 7), incident logs, and meeting records.

> *July 2024: We are changing the model again. We had a fine-tuned model; it was unstable, so we switched to one built from scratch; that has its own problems, so we consider switching back. I (A2) cannot finalize the interface while the underlying thing keeps changing, and I start to wonder whether the design was ever the problem. I (A1) keep moving between architectures and training runs, each one promising to be the fix, yet none clearly better. Weeks go by inside this loop.*

Model tuning repeatedly displaced interaction work. Architecture changes, sweeps, and transfer-learning choices came with long feedback loops and a standing temptation to treat a better model as the route to a better interaction, even when the real problem was timing, ambiguity, or a mapping [4]. A1's account of fine-tuning I3D matched this: gradual unfreezing helped, but the process was unstable and slow. A2 described the effect on the other side of the project as "going in circles" and as second-guessing whether the design was right. The tractable work of tuning quietly crowded out the harder, less measurable work of interaction design. **Implication:** Designers need explicit stopping rules and a hybrid path to fall back on to keep exploration alive when ML iteration starts to set the pace.

5.2.6. Theme 6: Recognition Error Reshapes Interpretation of the Data

This theme is grounded in 11 meaning units across diaries (A1: 6, A2: 5), incident logs, and meeting records.

> *August 2024: I (A2) was using the prototype the way it is meant to be used: looking into the low-birth-weight cases, the tail of the distribution*

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

where the clinically important records sit [82]. I make the gesture to filter down to that group. The system reads it wrong and shows me something else. For a second, I take the wrong view at face value and start to read a pattern into it. Then I realize the view is an artifact of the error, not the data. The mistake did not just fail a command. It briefly changed what I thought the data was telling me.

This is the theme the others lead to, and the one we did not set out to study. Supporting data analysis was not the prototype's goal; the goal was to test whether hand actions could drive the visualization. But testing that meant using the prototype for its analytic tasks ourselves, and in that use, we found that a recognition error did not stop at the interface. It changed our own reading of the data. A misrecognition while filtering or zooming not only failed to act; it introduced a doubt about whether a pattern was in the data or in the interaction. A2's reflections name this directly: "I underestimated how hard it would be to align gestures with how someone actually reasons about the data" and "every failed recognition risked pulling us away from what the numbers were telling us". The literature on trust and analytic work explains why such moments matter: trust, once disturbed, is slow to rebuild, and analytic reasoning depends on a stable coupling between a person and their view of the data [65, 66]. We registered this as designers, hesitation and lacking confidence, before any user study. This is exactly the early, interior signal a first-person method can catch. **Implication:** For prototypes whose task is analytic, ML uncertainty affected not only how the interaction felt but what we, its designers, could read from the data through it. The constraint that mattered was temporal alignment rather than offline accuracy. We claim this only for our own use; whether it holds for end users is the study this account is written to motivate.

5.3. Cross-map of Build Incidents to Themes

To keep the analysis auditable, Table 5 links recurring development situations from Stage 1 to the themes of Stage 2 and records, for each, what changed about what we could learn.

6. Discussion

We set out to understand what designers learn from prototyping with ML in the loop. Our prototype using naturalistic hand interactions gave us an answer in stages, and we experienced each stage as a struggle before we

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

Table 5: Audit cross-map linking build incidents (Stage 1) to analytic themes (Stage 2).

Design situation	**Stage 2 theme**	**Epistemic shift (what changed about "what we could learn")**
Manual scaffolding (a designer triggering by hand) was intermittently reintroduced during autonomous testing (e.g., confirmations, fallback triggers, smoothing) to keep exploration viable.	**Theme 1: Where Interpretation Sits Changes What "Working" Means**	From judging *interaction plausibility* to specifying *conditions of viability* (when autonomy is acceptable vs. when control must be reintroduced).
Disagreements and rework around labeling/segmentation and class boundaries (e.g., pruning/merging classes for learnability; redefining what "counts" as an action).	**Theme 2: Gesture and Dataset Work Is Interaction Design**	From exploring mappings to formalizing *action definitions* that determine the interaction vocabulary the prototype can legitimately test.
Real-time tests where latency and windowing broke the coupling between hand movement and system response (even when offline metrics looked acceptable).	**Theme 3: Timing, Not Accuracy, Governs Thinking-Through-Action**	From using accuracy as proxy evidence to treating *temporal alignment* as the primary validity constraint for interaction evidence.
Prototype progress became gated by pipeline readiness (data cleaning, feature extraction stability, runtime/inference constraints, integration glue).	**Theme 4: Labour Migrates into Pipelines and Infrastructure**	From iterating on interface ideas to maintaining *infrastructural conditions* required to generate trustworthy interaction evidence.
Model iteration cycles (fine-tuning, architecture changes, transfer learning) extended feedback loops and displaced interface experimentation.	**Theme 5: Iteration Loops Can Stall Exploration**	From rapid hypothesis testing to *long-cycle* evidence production; prompted explicit stopping rules and hybrid scaffolding to preserve exploration.
Visualization exploration tasks (e.g., inspecting tails, filtering, comparing groups) were disrupted by misrecognition and false triggers.	**Theme 6: Recognition Error Disturbed Our Own Reading of the Data**	From evaluating recognition in isolation to evaluating errors by their analytic consequence: how misrecognition disturbed *our own* reading of the data during use.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

understood it as a finding. Preparing the dataset taught us that deciding what counts as a gesture is interaction design (Theme 2). Wiring the model into the interface taught us that timing, not accuracy, decides whether the interaction survives (Theme 3). Keeping the system running taught us that our labour had moved into pipelines (Theme 4) and that model tuning could swallow design time (Theme 5). Moving between manual and autonomous interpretation taught us that each arrangement answers different questions (Theme 1). And using the prototype for its analytic tasks revealed a consequence we had not planned to study: recognition failures affected our own reading of the data (Theme 6). This section discusses these findings in relation to the literature, explains why ML still struggles with naturalistic hand actions, and draws out what follows for prototypes whose task is helping someone reason with data.

Mostly, the themes tell one story. Table 5 gives the incident-level trail behind this reading. The five conditions did not remain separate difficulties. Because testing the prototype meant using it for its analytic tasks ourselves, they converged as a mistimed or mistaken interpretation did not stop at the interface but reached our own conclusions about the data (Theme 6). The practical upshot is that the validity of such a prototype is temporal and interpretive rather than metric. A recognizer that scores well offline can still be the wrong evidence for whether an analytic interaction can be trusted, and a prototype that feels fluid under manual scaffolding can conceal exactly the failures that matter most once a person is using it to think.

6.1. RQ1: what can ML prototypes reveal about the feasibility, timing, and robustness of interpreting hand interactions?

Our build also documents why recognizing naturalistic hand actions remains hard. The root cause is that these actions carry design intent while being continuous, ambiguous, and settled only in context [17, 75, 74]. A grab, a point, and an inspection can look alike frame by frame; what separates them is what the person means to do. Current recognition pipelines cannot read that intent. They can only classify a discretized vocabulary, and every step of building that vocabulary cost us expressiveness. We felt this concretely. Mapping spontaneous movement onto fixed classes introduced inconsistencies that made recognition unreliable, and confusable classes pushed us to merge actions that are not the same to a hand (Theme 2); the per-class results of the fine-tuned I3D model (Figure 6) show exactly this pattern, with the ambiguous, intent-carrying actions scoring lowest. The 70-frame window that stabilized classification also delayed every response, so fast actions blurred

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

and trust fell (Theme 3). Fine-tuning a pretrained video model was unstable on our small dataset, and the tuning loop displaced interface work (Theme 5). None of these struggles is unique to us. Prior work reports the same pressures: hand actions resist rigid specification [50, 51], separating intent from incidental movement is open [52, 53], and human-AI interaction is distinctively difficult to design precisely because the capabilities and failures of the model are hard to anticipate [12, 10, 60]. What our first-person account adds is the shape these limits take inside a real build: they arrive as daily design decisions, each one trading a little of the interaction's naturalness for a little of the model's reliability.

Why insist on naturalistic hand actions for a data visualization at all? Because the hands are part of how people think about spatial and quantitative structure. Gesture externalizes ideas, scaffolds problem-solving, and coordinates spatial understanding [36, 37, 38, 41, 39]. A person who filters a group with a grab, or zooms into a distribution's tail with their fingers, is not entering commands. They are reasoning with their hands, and the interaction is meant to keep that reasoning moving. Our test case makes the stakes concrete: neonatal birth weight involves skewed distributions and rare but clinically consequential cases [83, 82], so exploring it means inspecting tails and comparing groups, exactly the operations our hand actions were designed to drive.

This is also why the system's failures cost more than a failed command. Reasoning with data depends on a stable coupling between the person, their action, and the view of the data [54, 65]. When recognition is late or wrong, that coupling breaks. In our own use, the break did not stop at the interface: a wrong view briefly read as a real pattern, and trust, once disturbed, was slow to return [66]. The practical lesson for anyone building hand-driven analytic systems is direct. If the system cannot deliver the visualization task reliably, in time, and on the intended action, it does not add friction. It subtracts from the reasoning the interaction was built to support.

6.2. RQ2: how these insights change when using manual vs. ML-based hand-interaction interpretation

Longstanding accounts frame prototyping as inquiry rather than implementation as a way of learning through the resistance of one's material [1, 2, 3, 4]. Our study specifies what changes in that learning when the material is a trained ML recognizer. In manual scaffolding, interpretive authority stays with the designer, and ambiguity can be repaired through judgment. In

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

autonomous recognition, interpretation moves to a probabilistic system, so uncertainty, latency, and misclassification become part of the interaction's texture rather than problems handled off-stage [9, 10, 64]. The important point is not that ML lowers or raises fidelity. It redistributes what the prototype can reveal: autonomy exposes failure modes that manual scaffolding conceals (feasibility, timing, robustness), while concealing what manual scaffolding makes plain (whether a breakdown belongs to the interaction or to the recognizer). Our hybrid arrangements made this trade legible, because moving deliberately between arrangements let us choose which question to make answerable at a given moment. On this account, the prototyping choice is not a step toward a finished system. It is an epistemic decision about what the designers will be able to learn.

Moreover, section 2 reviewed how past research characterizes working with learning systems: authority shifts to the model, labour is redistributed, and human-AI interaction becomes distinctively hard to design [9, 10, 11, 12]. Our account adds a consequence this literature leaves open. We are careful about its scope. Our prototype was not built to support data analysis; it was built to test whether naturalistic hand actions could drive a visualization. But because testing it meant using it for analytic tasks ourselves, we experienced something the past literature does not describe: when recognition failed, the error did not stop at the interaction layer. It propagated into our own interpretation of the data. That reframes the design target for prototypes like ours. The operative question is no longer only "can the model recognize the gesture?". It becomes "does the timing of recognition preserve the task the interaction is meant to serve?" It follows that the constraint governing these prototypes is temporal alignment rather than classification accuracy, a criterion that offline metrics, and much current ML tooling, do not capture [66, 65]. We offer this as the paper's primary conceptual contribution, scoped to designers' own use: whether the same effect holds for end users is an open question that invites, rather than forecloses, user-facing validation (Section 7).

6.3. What CAE Made Visible

Answering RQ3, this subsection names what the method itself made visible that a third-person evaluation would likely have missed.

Methodologically, CAE gave us access to dimensions of ML-driven design work that third-person evaluation tends to miss [26, 72, 33]. We could watch trust erode before any metric declined, see frustration function as a diagnostic signal rather than as noise, and trace how our disagreements shaped gesture

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

definitions, dataset scope, and system boundaries. These were not asides to the real work; they steered design decisions and iteration paths. Rather than treating our subjectivity as bias to be controlled, CAE let us render the difference between our roles analytically productive. A1's "model failure" and A2's "interaction breakdown" turned from a nuisance into evidence about how interpretation, delegation, and loss of control are negotiated as ML enters the loop. Paired with RTA [69], it produced an auditable account of how these insights formed over time, which is what makes a first-person account usable to others rather than personal. For partially autonomous systems, where the most consequential breakdowns are intermittent, contextual, and hard to reproduce, this kind of from-within account is not a lesser substitute for evaluation but a complementary source of design knowledge.

6.4. Implications for Designers

Our findings suggest four commitments for designers and HCI researchers building ML-driven prototypes for XR systems. Each comes from a specific moment in our design process.

- **Treat manual, autonomous, and hybrid interpretation as different instruments, not stages** Each makes different things learnable: manual scaffolding supports expressive exploration; autonomy exposes feasibility, error, and timing; hybrids let you move between the two without committing prematurely. We only understood the feasibility of our own gestures after the hand-off to the model broke them (Theme 1). Sequence arrangements by what you need to learn, not by assumed fidelity. In practice: in a Unity or Unreal project, keep a controller- or keyboard-triggered path for every action alongside the recognizer, switchable per build flag, so a manually scaffolded pass is always one toggle away.

- **Recognize gesture and dataset work as interaction design** Segmentation, categorization, and annotation quietly author the interaction vocabulary. Our merged classes and pruned actions were design decisions made at the labelling table (Theme 2). Make them explicit, and revisit them as design decisions rather than data chores. In practice: version the label set the way you version code, and when two XR gestures confuse the recognizer, record the merge as a design decision with its interaction cost, not only as an accuracy gain.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

- **Evaluate temporal validity, not only accuracy** In our use, the 70-frame delay undermined trust more than the occasional misclassification did (Theme 3). Judge an ML-driven prototype by the task it can sustain, not only by the gestures it can classify. In practice: log per-action latency from hand movement to visible response in the headset, and treat any perceptible delay as a design defect to triage, not a metric to average away; in our build, the 70-frame window was the defect, not the error rate.

- **Bound ML iteration to protect exploration** Without stopping rules, model tuning eclipses interface design; our fine-tuned versus scratch loop consumed weeks (Theme 5). Set explicit exit criteria, and keep a hybrid path available so exploration continues while the model catches up. In practice: timebox tuning sprints before they begin, and require that every sprint ends with the prototype runnable end-to-end, so interface learning never waits on model learning.

Moreover, these commitments describe a staged, reflective approach to prototyping with ML: rather than asking whether ML can replace earlier techniques, treat it as one arrangement within a broader repertoire, deployed deliberately and in dialogue with the task the interaction is meant to serve.

Finally, the pairing at the centre of our prototype points beyond it. We would not ask people to learn a gesture set; we let the movements they already make while reasoning about objects in space, on the pieces in their hands, stand in for analytic operations that resemble them (Section 4.2). For spatial user interfaces in XR this suggests a design route that differs from command gestures. Headsets already track hands and objects [8, 7], so the pieces that our webcam prototype needs on the desk can become tracked physical props or virtual objects in the scene, and the same vocabulary of spatial hand actions can drive a visualization that surrounds the person rather than sits on a screen. Two consequences of our account carry over. First, the recognizer's vocabulary is tied to the material it was learned on, so a system that changes the objects a person handles changes what can be recognized; the objects are part of the interface, not incidental to it. Second, an action that issues no command, our Mental Image marker, is still informative: it tells the interface that spatial reasoning is under way. A spatial interface that reads such markers could time its guidance to the person's thinking rather than to a click. We offer these as directions.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

7. Limitations

Our account rests on two designer-researchers with deliberately differentiated roles. We have argued that this difference is part of our instrument. But it also bounds the diversity of practice, discipline, and infrastructure the study represents. Our findings are best read as transferable propositions to be tested with other teams, domains, and kinds of ML, rather than as generalizations. Most importantly for our central claim, our evidence that recognition error reshapes interpretation is our own lived experience: we did not run a user study, so this consequence is demonstrated in design practice and remains to be established with end users. A study that deliberately induces recognition errors during analytic tasks and measures their effect on interpretation, trust, and decision-making is the natural and necessary next step, and the one this paper is written to motivate. We also note the cost of the method itself. Reflecting and writing while still designing were demanding, and at times they competed with the design work for attention. A2 recorded that documenting the process, while useful for understanding our decisions, "slowed me down" and was "exhausting". We managed this by keeping short notes during the work and reserving longer reflection for milestones, but we report it plainly, because it shapes what a first-person account of live design work can and cannot capture [71]. Finally, as with all autoethnography, our account is interpretive and situated by design [14, 13, 67]; its value lies not in detachment but in surfacing, from inside practice, the questions and constraints that third-person evaluation tends to miss. We have tried to make that situatedness auditable, so that readers can judge our interpretations against the trail that produced them.

8. Conclusion

When designers build naturalistic hand-based interaction with ML in the loop, the decisive change is not to fidelity but to knowing. Through an eight-month collaborative autoethnography of our own practice, we found that in tasks where the interaction itself is a way of reasoning about data, recognition error and latency extends beyond the interface and into interpretation itself. Whether such a prototype yields trustworthy evidence turns on temporal alignment rather than offline accuracy. Around that central consequence lie the conditions that produce it: where interpretation sits, gesture-and-dataset work as design, and the pull of pipelines and iteration loops. We therefore

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

argue for staged, hybrid prototyping that keeps interpretation legible while confronting ML's temporal and probabilistic limits, and for judging these prototypes by the reasoning they sustain, not only the gestures they classify. More broadly, we offer prototyping with ML in the loop as a site of genuine epistemic work, and collaborative autoethnography as a way to study it honestly from within.

CRediT authorship contribution statement

Prashant Rawat: Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Sussan T. Anukem:** Methodology, Software, Investigation, Data curation, Formal analysis, Writing – original draft, Visualization. **Rita Orji:** Supervision, Writing - review & editing. **Ga Wu:** Resources, Writing - review & editing. **Francisco Raul Ortega:** Resources, Writing - review & editing. **Lizbeth Escobedo Bravo:** Supervision, Methodology, Writing - review & editing. **Mayra Donaji Barrera Machuca:** Supervision, Conceptualization, Resources, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The development log, reflection prompts, trained recognition model, and inference notebook are provided as supplementary material. The diary corpus contains the authors' personal reflections and is available from the corresponding author on reasonable request.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

References

[1] S. Houde, C. Hill, What do prototypes prototype?, in: Handbook of human-computer interaction, Elsevier, 1997, pp. 367–381.

[2] Y.-K. Lim, E. Stolterman, J. Tenenberg, The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas, ACM Transactions on Computer-Human Interaction (TOCHI) 15 (2) (2008) 1–27.

[3] L. Kimbell, Rethinking design thinking: Part ii, Design and Culture 4 (2) (2012) 129–148.

[4] D. A. Schön, The reflective practitioner: How professionals think in action, Routledge, 2017.

[5] Unity Technologies, Unity real-time development platform (2024). URL https://unity.com

[6] Epic Games, Unreal engine (2024). URL https://www.unrealengine.com

[7] Meta Platforms, Meta XR interaction SDK documentation (2024). URL https://developers.meta.com/horizon/documentation/unity/unity-isdk-interaction-sdk-overview/

[8] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, et al., Megatrack: monochrome egocentric articulated hand-tracking for virtual reality, ACM Transactions on Graphics (ToG) 39 (4) (2020) 87–1.

[9] S. Amershi, M. Cakmak, W. B. Knox, T. Kulesza, Power to the people: The role of humans in interactive machine learning, AI magazine 35 (4) (2014) 105–120.

[10] G. Dove, K. Halskov, J. Forlizzi, J. Zimmerman, Ux design innovation: Challenges for working with machine learning as a design material, in: Proceedings of the 2017 chi conference on human factors in computing systems, 2017, pp. 278–288.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[11] Q. Yang, N. Banovic, J. Zimmerman, Mapping machine learning advances from hci research to reveal starting places for design innovation, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–11.

[12] Q. Yang, A. Steinfeld, C. Rosé, J. Zimmerman, Re-examining whether, why, and how human-ai interaction is uniquely difficult to design, in: Proceedings of the 2020 chi conference on human factors in computing systems, 2020, pp. 1–13.

[13] H. Chang, F. Ngunjiri, K.-A. C. Hernandez, Collaborative autoethnography, Routledge, 2016.

[14] C. Ellis, T. E. Adams, A. P. Bochner, Autoethnography: an overview, Historical social research/Historische sozialforschung (2011) 273–290.

[15] S. Schlögl, G. Doherty, S. Luz, Wizard of oz experimentation for language technology applications: Challenges and tools, Interacting with Computers 27 (6) (2015) 592–615.

[16] N. O. Bernsen, H. Dybkjær, L. Dybkjær, Wizard of oz prototyping: How and when, Proc. CCI Working Papers Cognit. Sci./HCI, Roskilde, Denmark (1994).

[17] L. A. Suchman, Plans and situated actions: The problem of human-machine communication, Cambridge university press, 1987.

[18] P. Green, L. Wei-Haas, The rapid development of user interfaces: Experience with the wizard of oz method, in: Proceedings of the human factors society annual meeting, Vol. 29, SAGE Publications Sage CA: Los Angeles, CA, 1985, pp. 470–474.

[19] S. Dow, B. MacIntyre, J. Lee, C. Oezbek, J. D. Bolter, M. Gandy, Wizard of oz support throughout an iterative design process, IEEE Pervasive Computing 4 (4) (2005) 18–26.

[20] N. T. T. Mai, T. T. T. Hai, N. V. Son, Wizard of oz for designing hand gesture vocabulary in human-robot interaction, in: 2011 Third International Conference on Knowledge and Systems Engineering, IEEE, 2011, pp. 232–238.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[21] B. Buxton, Sketching user experiences: getting the design right and the right design, Morgan kaufmann, 2010.

[22] P. Dourish, Implications for design, in: Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006, pp. 541–550.

[23] J. Clifford, On ethnographic authority, Representations (2) (1983) 118–146.

[24] T. Salvador, G. Bell, K. Anderson, Design ethnography, Design Management Journal (Former Series) 10 (4) (1999) 35–41.

[25] R. Bentley, J. A. Hughes, D. Randall, T. Rodden, P. Sawyer, D. Shapiro, I. Sommerville, Ethnographically-informed systems design for air traffic control, in: Proceedings of the 1992 ACM conference on Computer-supported cooperative work, 1992, pp. 123–129.

[26] A. Lucero, Living without a mobile phone: an autoethnography, in: Proceedings of the 2018 Designing Interactive Systems Conference, 2018, pp. 765–776.

[27] A. Desjardins, A. Ball, Revealing tensions in autobiographical design in hci, in: proceedings of the 2018 designing interactive systems conference, 2018, pp. 753–764.

[28] K. Spiel, " why are they all obsessed with gender?"—(non) binary navigations through technological infrastructures, in: Proceedings of the 2021 ACM Designing Interactive Systems Conference, 2021, pp. 478–494.

[29] C. Neustaedter, P. Sengers, Autobiographical design in hci research: designing and learning through use-it-yourself, in: Proceedings of the Designing Interactive Systems Conference, 2012, pp. 514–523.

[30] R. D. Sawyer, D. Lund, Duoethnography: Dialogic methods for social, health, and educational research, Vol. 7, Left Coast Press, 2012.

[31] J. Zimmerman, J. Forlizzi, S. Evenson, Research through design as a method for interaction design research in hci, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 2007, pp. 493–502.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[32] W. Gaver, What should we expect from research through design?, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2012, pp. 937–946.

[33] A. Rapp, In search for design elements: a new perspective for employing ethnography in human-computer interaction design research, International Journal of Human–Computer Interaction 37 (8) (2021) 783–802.

[34] G. Pritchard, J. Vines, P. Briggs, L. Thomas, P. Olivier, Digitally driven: how location based services impact the work practices of london bus drivers, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014, pp. 3617–3626.

[35] X. Xiao, H. Ishii, Inspect, embody, invent: a design framework for music learning and beyond, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 5397–5408.

[36] A. Kendon, Gesture: Visible action as utterance, Cambridge University Press, 2004.

[37] B. Tversky, Mind in motion: How action shapes thought, Hachette UK, 2019.

[38] S. Chandrasegaran, D. Ramanujan, N. Elmqvist, How do sketching and non-sketching actions convey design intent?, in: Proceedings of the 2018 Designing Interactive Systems Conference, 2018, pp. 373–385.

[39] M. Studdert-Kennedy, Hand and mind: What gestures reveal about thought., Language and Speech 37 (2) (1994) 203–209.

[40] J. B. Haviland, Pointing, gesture spaces, and mental maps, Language and gesture 2 (2000) 13.

[41] M. A. Novack, S. Goldin-Meadow, Gesture as representational action: A paper about function, Psychonomic Bulletin & Review 24 (2017) 652–665.

[42] W.-M. Roth, D. Lawless, Scientific investigations, metaphorical gestures, and the emergence of abstract scientific concepts, Learning and instruction 12 (3) (2002) 285–304.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[43] Y. Yin, R. Davis, Toward natural interaction in the real world: Real-time gesture recognition, in: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, 2010, pp. 1–8.

[44] H. Sampson, D. Kelly, B. C. Wünsche, R. Amor, A hand gesture set for navigating and interacting with 3d virtual environments, in: 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ), IEEE, 2018, pp. 1–6.

[45] S. Wu, T. Gable, K. May, Y. M. Choi, B. N. Walker, Comparison of surface gestures and air gestures for in-vehicle menu navigation, Archives of Design Research 29 (4) (2016) 65–80.

[46] J. Aliprantis, M. Konstantakis, R. Nikopoulou, P. Mylonas, G. Caridakis, Natural interaction in augmented reality context., in: VIPERC@ IRCDL, 2019, pp. 50–61.

[47] W. Lin, L. Du, C. Harris-Adamson, A. Barr, D. Rempel, Design of hand gestures for manipulating objects in virtual reality, in: Human-Computer Interaction. User Interface Design, Development and Multimodality: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 19, Springer, 2017, pp. 584–592.

[48] J. L. Rubio-Tamayo, M. Gertrudix Barrio, F. García García, Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation, Multimodal technologies and interaction 1 (4) (2017) 21.

[49] V. Chang, R. O. Eniola, L. Golightly, Q. A. Xu, An exploration into human–computer interaction: Hand gesture recognition management in a challenging environment, SN Computer Science 4 (5) (2023) 441.

[50] V. A. Shanthakumar, C. Peng, J. Hansberger, L. Cao, S. Meacham, V. Blakely, Design and evaluation of a hand gesture recognition approach for real-time interactions, Multimedia Tools and Applications 79 (25) (2020) 17707–17730.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[51] X. Lou, X. A. Li, P. Hansen, P. Du, Hand-adaptive user interface: improved gestural interaction in virtual reality, Virtual Reality 25 (2021) 367–382.

[52] H. Wu, W. Luo, N. Pan, S. Nan, Y. Deng, S. Fu, L. Yang, Understanding freehand gestures: a study of freehand gestural interaction for immersive vr shopping applications, Human-centric Computing and Information Sciences 9 (2019) 1–26.

[53] A. Masurovsky, P. Chojecki, D. Runde, M. Lafci, D. Przewozny, M. Gaebler, Controller-free hand tracking for grab-and-place tasks in immersive virtual reality: Design elements and their empirical study. multimodal technologies and interaction 4, 4 (2020) (2020).

[54] P. Dourish, What we talk about when we talk about context, Personal and ubiquitous computing 8 (2004) 19–30.

[55] W. Hou, D. Xia, H. Jung, Video road vehicle detection and tracking based on opencv, in: 2020 international conference on information science and education (ICISE-IE), IEEE, 2020, pp. 315–318.

[56] D. Schleicher, P. Jones, O. Kachur, Bodystorming as embodied designing, interactions 17 (6) (2010) 47–51.

[57] S. Greenberg, S. Carpendale, N. Marquardt, B. Buxton, Sketching user experiences: The workbook, Elsevier, 2011.

[58] J. M. Vega-Cebrián, A. Lindegren, M. Gamboa, A. Tajadura-Jiménez, Y. Fernaeus, E. Márquez Segura, Embodied ideation, toolkits, and sketching, in: Proceedings of the Nineteenth International Conference on Tangible, Embedded, and Embodied Interaction, 2025, pp. 1–6.

[59] S. Connell, P.-Y. Kuo, L. Liu, A. M. Piper, A wizard-of-oz elicitation study examining child-defined gestures with a whole-body interface, in: Proceedings of the 12th International Conference on Interaction Design and Children, 2013, pp. 277–280.

[60] E. Mitchell, L. Mamykina, From the curtain to kansas: Conducting wizard-of-oz studies in the wild, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–6.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[61] G. Zhu, L. Zhang, P. Shen, J. Song, Multimodal gesture recognition using 3-d convolution and convolutional lstm, Ieee Access 5 (2017) 4517–4524.

[62] C. Jian, J. Li, M. Zhang, Lstm-based dynamic probability continuous hand gesture trajectory recognition, IET Image Processing 13 (12) (2019) 2314–2320.

[63] M. Maghoumi, J. J. LaViola, Deepgru: Deep gesture recognition utility, in: Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14, Springer, 2019, pp. 16–31.

[64] Q. Yang, A. Scuito, J. Zimmerman, J. Forlizzi, A. Steinfeld, Investigating how experienced ux designers effectively work with machine learning, in: Proceedings of the 2018 designing interactive systems conference, 2018, pp. 585–596.

[65] G. Klein, B. Moon, R. R. Hoffman, Making sense of sensemaking 1: Alternative perspectives, IEEE intelligent systems 21 (4) (2006) 70–73.

[66] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (1) (2004) 50–80.

[67] A. Rapp, Autoethnography in human-computer interaction: Theory and practice, New directions in third wave human-computer interaction: Volume 2-Methodologies (2018) 25–42.

[68] P. Geist-Martin, L. Gates, L. Wiering, E. Kirby, R. Houston, A. Lilly, J. Moreno, Exemplifying collaborative autoethnographic practice via shared stories of mothering, Journal of Research Practice 6 (1) (2010) M8–M8.

[69] V. Braun, V. Clarke, One size fits all? what counts as quality practice in (reflexive) thematic analysis?, Qualitative research in psychology 18 (3) (2021) 328–352.

[70] T. Winograd, On understanding computers and cognition: A new foundation for design, Artificial intelligence 31 (2) (1987) 250–261.

[71] J. C. Lapadat, Ethics in autoethnography and collaborative autoethnography, Qualitative Inquiry 23 (8) (2017) 589–603. doi:10.1177/1077800417704462.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[72] A. Desjardins, R. Wakkary, Living in a prototype: A reconfigured space, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 5274–5285.

[73] P. Rawat, M. D. Barrera Machuca, Spatial hand actions: Exploring the hand actions used to represent spatial thinking for 3d assembling tasks, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. doi:10.1145/3706598.3713424.
URL https://doi.org/10.1145/3706598.3713424

[74] B. Tversky, A. Kessell, Thinking in action, Pragmatics & Cognition 22 (2) (2014) 206–223.

[75] M. W. Alibali, Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information, Spatial cognition and computation 5 (4) (2005) 307–331.

[76] J. S. Yi, Y. ah Kang, J. Stasko, J. A. Jacko, Toward a deeper understanding of the role of interaction in information visualization, IEEE transactions on visualization and computer graphics 13 (6) (2007) 1224–1231.

[77] M. Brehmer, T. Munzner, A multi-level typology of abstract visualization tasks, IEEE transactions on visualization and computer graphics 19 (12) (2013) 2376–2385.

[78] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, et al., Mediapipe: A framework for building perception pipelines, arXiv preprint arXiv:1906.08172 (2019).

[79] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: Lstm cells and network architectures, Neural computation 31 (7) (2019) 1235–1270.

[80] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495

[81] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, K. Alahari, Charades-ego: A large-scale dataset of paired third and first person videos, arXiv preprint arXiv:1804.09626 (2018).

[82] C. L. Cutland, E. M. Lackritz, T. Mallett-Moore, A. Bardají, R. Chandrasekaran, C. Lahariya, M. I. Nisar, M. D. Tapia, J. Pathirana, S. Kochhar, et al., Low birth weight: Case definition & guidelines for data collection, analysis, and presentation of maternal immunization safety data, Vaccine 35 (48) (2017) 6492–6500.

[83] H. Blencowe, J. Krasevec, M. De Onis, R. E. Black, X. An, G. A. Stevens, E. Borghi, C. Hayashi, D. Estevez, L. Cegolon, et al., National, regional, and worldwide estimates of low birthweight in 2015, with trends from 2000: a systematic analysis, The Lancet global health 7 (7) (2019) e849–e860.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7442495