

Generative Proxy: Synthesizing Proxy-Based Interfaces for Real-World Interaction Across AR Glasses

Xianhao Carton Liu
liu03008@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

William Chastek
chast068@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

Eric J Gonzalez
ejgonz@google.com
Google
Seattle, Washington, USA

Mar Gonzalez-Franco
margon@google.com
Google
Seattle, Washington, USA

Chen Zhu-Tian
ztchen@umn.edu
University of Minnesota
Minneapolis, Minnesota, USA

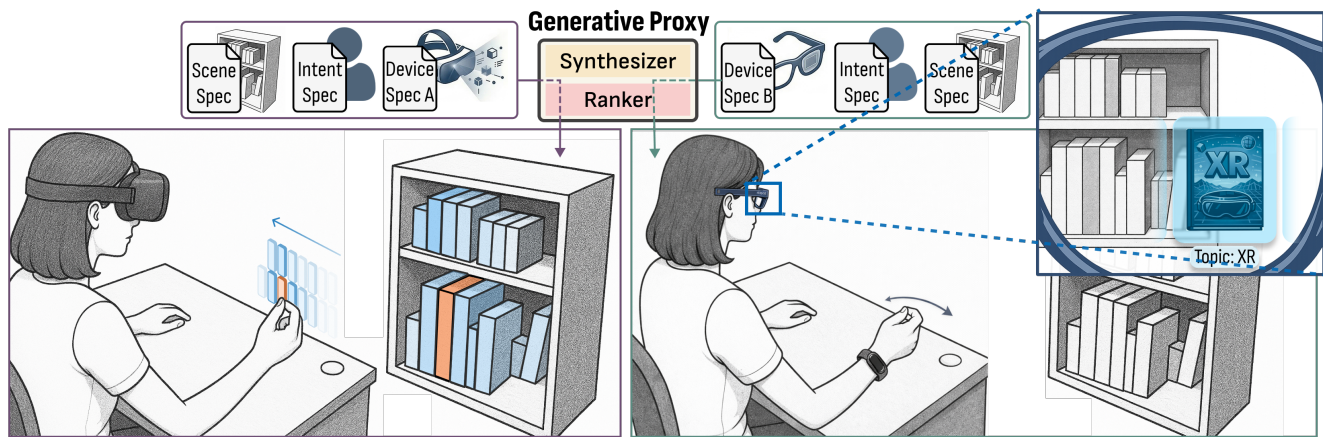


Figure 1: Generated proxy UIs for immersive AR headsets (left, adapted from [38]) and lightweight AR glasses (right). Generative Proxy accounts for diverse scenes, user intents, and device profiles to produce appropriate proxy UIs. As shown on right, proxy UIs enable interaction with real-world objects on lightweight AR glasses, despite the lack of ray casting or pointing input.

Abstract

Interacting with real-world objects in AR is difficult, especially when targets are distant, cluttered, or occluded. These challenges are amplified on emerging lightweight AR glasses, which often lack binocular or large field of view on display, but also continuous inputs, such as hand or eye tracking. Proxy-based interfaces offer an alternative by allowing users to interact with virtual abstractions of physical objects that can be repositioned, reorganized, and adapted to the task and device. However, designing such interfaces is currently manual and highly device-specific. We present Generative Proxy, a method for automatically generating proxy-based interfaces from three specifications: scene, intent, and device capabilities. We formulate generation as a constrained synthesis problem that first produces valid interfaces for the target device and task, then ranks candidates using semantic and articulatory distance inspired

by direct manipulation theory. We demonstrate Generative Proxy across diverse scenes, device profiles, and user intents. Expert evaluation shows initial evidence that generated proxy UIs are useful and usable, highlighting proxy-based abstraction as a promising interaction paradigm for future AR glasses.

Keywords

AR/MR/XR, Generative UI, Gestural Interaction, AR Glasses

ACM Reference Format:

Xianhao Carton Liu, William Chastek, Eric J Gonzalez, Mar Gonzalez-Franco, and Chen Zhu-Tian. 2026. Generative Proxy: Synthesizing Proxy-Based Interfaces for Real-World Interaction Across AR Glasses. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3830398.3830482>



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
UIST '26, Detroit, MI, USA

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2856-3/2026/11
<https://doi.org/10.1145/3830398.3830482>

1 Introduction

Recent advances in glasses-like wearable computing suggest a shift from bulky immersive headsets toward lighter, more wearable augmented reality (AR) form factors. Companies including Google, Meta, and Snap are actively developing glasses-based platforms, reflecting growing interest in AR experiences embedded in everyday

interaction with the physical world [52]. As these devices become more practical and socially acceptable, a central challenge is how best to facilitate interaction with the real-world around the user.

A common approach in AR is to interact with physical objects directly using ray-casting, gaze, or other remote targeting techniques [17, 48]. However, direct interaction is constrained by the physical scene itself: objects may be distant, cluttered, occluded, or awkward to access [63]. Moreover, it is often a poor fit for lightweight AR glasses, whose input capabilities may be limited to voice or discrete gestures rather than continuous mid-air pointing [33].

Proxy-based interfaces [38] offer an complementary paradigm to direct interactions. Instead of selecting or manipulating physical objects directly, users interact with virtual surrogates that represent them. Because these surrogates are decoupled from the objects' physical arrangement, they can be repositioned, resized, and reorganized to better support the user's task. For example, rather than requiring a user to target multiple books, a proxy can group relevant books into a single representation that is easier to inspect and select. Proxy-based interfaces are therefore especially useful when targets are distant, cluttered, occluded, visually similar, or otherwise difficult to reference unambiguously.

However, proxy-based interfaces remain largely handcrafted and device-specific. AR devices vary widely in display and input capabilities, from spatially registered 3D content to 2D overlays, and from full hand tracking to gaze or discrete gestures [30, 39, 41, 45]. As a result, a proxy UI designed for spatial 3D presentation with continuous hand input (Figure 1 Left) and may not be appropriate for another device with only screen-space output and discrete gestures (Figure 1 Right). Manually redesigning and retuning such interfaces for every device therefore does not scale.

In this work, we ask: *how can proxy-based interfaces be generated automatically from the physical scene, the user's intent, and the target device capabilities?* We present Generative Proxy, a method for synthesizing proxy-based interfaces for real-world interaction across heterogeneous AR capability conditions. Given typed specifications of scene, intent, and device capability, Generative Proxy first synthesizes interface candidates that are valid for the target task and platform, then ranks them using direct-manipulation-inspired criteria that favor interfaces with lower semantic, articulatory, and evaluation distance [24]. This yields proxy interfaces that are both device-compatible and better aligned with the user's task.

We demonstrate Generative Proxy through a proof-of-concept implementation and three application scenarios, and further probe the pipeline through an internal ranking sanity check and an expert evaluation with experienced XR users ($N = 12$). Together, these studies provide initial evidence that proxy-based UIs can support interaction with real-world objects across a broader range of AR capability conditions, and that our method can generate plausible and usable proxy UIs across varying device capabilities, scenes, and intents. In summary, this paper makes the following contributions:

- **Design.** A formal specification framework for representing proxy interfaces, physical scenes, user intents, and device capabilities, together with a constrained synthesis-and-ranking pipeline for generating proxy interfaces in AR.

- **Prototype.** A proof-of-concept cross-device implementation and three application scenarios demonstrating how proxy-based UIs can be generated and applied across diverse AR capability profiles.
- **Evaluation.** Initial evidence from an internal ranking sanity check and an expert evaluation ($N=12$) probing the usefulness of generated proxy UIs, how device-specific variants are perceived, and where the current pipeline breaks down.

2 Related Work

This work concerns interacting with real-world objects in AR for digital tasks. We review prior work in three areas: interaction with real-world objects in AR, UI synthesis and generation, and contextually adaptive AR interfaces.

2.1 Interacting with Real-World Objects in AR

We use *interacting with real-world objects in AR* to describe systems that treat physical objects as referents for associated digital information or actions. We organize this literature through a hardware- and interaction-driven lens, since AR platforms have strongly shaped how such interactions are realized.

Handheld and screen-mediated interactions. Early AR systems typically accessed physical objects through the device itself, for example via fiducial markers [27] or handheld mobile screens [21]. Marker-based systems [1, 26] attached digital information and actions to instrumented referents through explicit tags, while mobile AR systems such as *RealitySketch* [53] supported object-linked content through touchscreen interaction. In these systems, the device remained the primary locus of interaction.

Head-worn embodied interactions. With head-worn AR, interaction shifted from the device toward embodied spatial input. Systems in this space use gaze, head motion, hand rays, gestures, voice, and direct touch-like interaction to specify physical targets and invoke digital actions. A recurring theme is multimodal disambiguation: *Pinpointing* [28] combines head motion and gaze for precise AR selection, *GazePointAR* [32] combines gaze, gesture, and speech for situated reference resolution, and *Uncertain Pointer* [58] visualizes referential ambiguity during AR target selection. A parallel line of work treats everyday objects as interfaces [10, 22, 44, 57, 68]. For example, *XR-Objects* [9] uses multimodal large language models to let real objects expose context-sensitive menus and actions without preregistration. Together, these systems move beyond device mediation toward embodied, object-grounded interaction.

Lightweight wearable and distributed interactions. Lightweight AR glasses introduce a different regime of constraints, including limited sensing fidelity, restricted feedback bandwidth, and reduced support for precise manual input [34]. To address these limitations, Ren et al. developed a peripheral AR display technique to facilitate real-world object selection on lightweight AR glasses [50]. More broadly, recent work increasingly relies on distributed sensing and AI-mediated inference to recover user intent from partial signals. These systems combine cues such as speech [43], coarse pointing [32], gaze [6], wrist- or ring-based input [5], and egocentric scene context [9] with learned models to infer the intended object or action. Examples include *SensibleAgent* [31] for context-aware

interaction and Squiggle [12] for inferring multi-object reference through implicit lasso input. We do not review the broader literature on wearable gesture and activity recognition [4, 62], which is adjacent but not central here.

Summary. Across these lines of work, interaction typically remains tied to directly referring to, touching, or pointing at the target object itself. Proxy-based interfaces redirect interaction from the physical referent to an abstract or synthesized proxy that preserves its digital semantics while reducing the need for direct access to the object. *RealityProxy* [38] is a prior instantiation of this broader interaction paradigm. Building on this foundation, our work asks how proxy interfaces can be synthesized across heterogeneous AR systems with varying sensing, ergonomic, and form-factor constraints.

2.2 User Interface Synthesis and Generation

The common premise of prior work on UI generation is that interfaces can be derived from representations more abstract than the final UI, such as task models, declarative specifications, or natural-language descriptions. They differ mainly in how much structure is provided in advance and how much control remains over the generation process.

Model-based user interfaces. Model-based UI (MBUI) approaches derive interfaces from abstract models such as task models, domain models, and mappings from abstract interactions to concrete widgets [46, 55, 65]. This separation supports systematic generation, retargeting, and consistency. More recent examples include ORC Layout [25], which formulates adaptive GUI layout with ORC constraints, and Scout [54], which supports mixed-initiative exploration of layout alternatives through high-level constraints. Yet, MBUI often requires substantial upfront modeling and tends to assume relatively static, developer-authored structures.

Specification-based generation. Specification-based generation can be seen as a more constrained form of model-based generation, centered on structured intermediate representations that can be interpreted reliably and sometimes manipulated directly. For example, Bespoke [60] synthesizes GUIs for command-line workflows from demonstrations, while Vega-Lite [51] provides a declarative grammar for visualization specification and systems such as DynaVis [59] build persistent editing interfaces on top of such representations. The main advantage is interpretability and control: explicit primitives and mappings make the generation space more reusable and consistent. The tradeoff is that expressiveness is bounded by the underlying specification language.

AI-based constraint-guided generation. Recent work has pushed UI generation toward direct synthesis from natural-language prompts and other high-level inputs using large language models. This lowers authoring barriers and broadens the range of tasks that can be addressed, but direct prompt-to-code generation remains difficult to control and revise [7]. Systems such as UICoder [64], LLM-assisted design support in Figma [15], Misty [40], and Squire [36] respond by reintroducing structure around AI generation through feedback, intermediate representations, and mixed-initiative workflows. Related work on malleable interfaces [3, 42] further treats AI as a

mechanism for continued interface evolution rather than only initial generation. Together, this literature suggests that AI expands the scope of UI generation, but benefits from structured intermediates, explicit constraints, and mixed-initiative interaction.

Building on these threads, we adopt a hybrid specification-and-AI approach: specification-based synthesis constrains the design space to valid candidates, while a theory-grounded LLM-based ranker prioritizes interfaces with lower semantic, articulatory, and evaluation distance.

2.3 Contextually Adaptive Interfaces in AR

Adaptive interfaces have long been central to AR because decisions about what to show, where to place it, and how it should appear depend on the user’s immediate physical context [18]. A large body of work has therefore focused on adaptation grounded in environmental geometry [13, 14, 47, 67]. For example, AdapTUI [20] formulates adaptation as an optimization problem that jointly considers geometric and human factors for tangible AR interfaces.

More recent work extends adaptation beyond geometry by incorporating semantics of the environment. Lindlbauer et al. [37] adapt placement and level of detail based on contexts and user cognitive load; Tahara et al. [56] use scene graphs to maintain consistency across locations; AdapTutAR [23] personalizes AR instructions to user characteristics; SemanticAdapt [8] uses computer vision to attach semantically relevant augmentations; and Zhu-Tian et al. [66] apply reinforcement learning to continuously adapt label layouts. Authoring tools such as CAPtutAR [61], AUIT [11], and ScalAR [49] further lower the barrier to building adaptive AR applications.

Our work shares this interest but differs in where adaptation enters the pipeline. Rather than adapting pre-authored interfaces online, we cast the problem as one of UI synthesis from device capability, user intent, and scene structure. The resulting interfaces could subsequently be combined with runtime adaptation techniques for further refinement of layout and placement.

3 Problem Formulation

Proxy interfaces has the potential to facilitate real-world interactions across heterogeneous AR capability conditions, each with different sensing, input, and output capabilities. However, manually redesigning a proxy interface for every device is therefore impractical and does not scale. This motivates the need for a method that can generate proxy interfaces automatically while remaining appropriate to the user, the environment, and the target device. From this problem setting, we derive two requirements:

R1: Generating Valid Interfaces from Scene, Intent and Device Capability. Whether a proxy interface is valid depends on three inputs: (1) *scene*, which specifies the available physical objects and their relevant attributes; (2) *intent*, which captures what the user is trying to do; and (3) *device capability*, which defines the target platform’s available input and output. A generated proxy interface is valid only if it is compatible with the device capabilities and supports the intended interaction over the given scene.

R2: Ranking Valid Interfaces by Directness of Interaction. Validity alone is insufficient. For a given scene, intent and device, many proxy interfaces may satisfy these basic constraints, yet differ

substantially in usability. The remaining problem is therefore not only to generate feasible interfaces, but also to determine which ones better preserve the feel of direct manipulation.

We frame this distinction using Hutchins, Hollan, and Norman’s account of direct manipulation [24], which relates perceived directness to the gulfs of execution and evaluation. In particular, the gulf of execution depends in part on *semantic distance*, or how closely the available actions correspond to the user’s intended task, and *articulatory distance*, or how easily those actions can be physically expressed. This perspective is especially relevant for proxy-based interfaces because they decouple interaction from the original physical arrangement of objects. That decoupling creates opportunities to reorganize interactive representations in ways that better match the user’s task and the device’s interaction constraints. For example, if a user wants to select all HCI books, a proxy could group them into a single nearby representation, reducing semantic distance by aligning the interface with the task structure and reducing articulatory distance by enabling a simple within-reach action.

Accordingly, beyond generating valid proxy interfaces, the problem also requires prioritizing those that reduce semantic and articulatory distance for the intended interaction.

4 Generative Proxy: Approach

We formulate proxy interface generation as a constrained synthesis problem. Given a scene specification, a user intent specification, and a device capability specification, the system synthesizes candidate interfaces and ranks feasible candidates using direct-manipulation-inspired distances. Figure 2 illustrates the pipeline through a running bookshelf example for partial display AR glasses.

4.1 Typed Specifications

We define a typed specification for the scene, user intent, device capabilities, and proxy interfaces. Together, these specifications make the synthesis space explicit and provide contracts for what is present in the environment, what the user wants to do, what

the device can support, and what counts as a valid interface. Below we summarize each specification and its design rationale; full definitions are provided in the supplementary materials.

4.1.1 Scene. We represent the Scene as a set of typed entities:

$$\text{Scene} := \{(\text{name}, \text{type}, \text{attributes}, \text{pos}, \text{parent}?)\}$$

Each entity records its identity, semantic type, attributes, approximate position, and optional parent relation, similar to the abstractions used in Reality Proxy [38]. In practice, this level of abstraction is sufficient for deciding which objects can be proxied together, which objects share meaningful structure, and how a generated interface should remain grounded in the real world. For example, the input SceneSpec in Figure 2 describes bookshelf entities with attributes such as topic and color.

4.1.2 User Intents. The user’s goal is described by an IntentSpec:

$$\text{IntentSpec} := (\text{kind}, \text{scope}, \text{by}?)$$

kind specifies the operation to be performed, scope specifies what it ranges over, and by optionally refines that operation, for example by grouping or filtering criterion. A key design decision is that IntentSpec captures *what* the user wants to accomplish without prescribing *how* the interface should realize it. This separation is essential for synthesis: the same high-level goal, such as inspecting distant objects or reorganizing a cluttered set, may be realized by very different proxy interfaces depending on the scene and the target device. In this work, we implemented five intents (i.e., *inspect*, *select*, *organize*, *filter*, and *relevel*) based on interaction taxonomies from prior work [30] and the operations supported in Reality Proxy [38], although other choices are possible. The example IntentSpec in Figure 2 requests inspect, select, and filter.

4.1.3 Device Capabilities. A CapabilitySpec captures the interaction resources available on a target device:

$$\text{CapabilitySpec} := (\{\text{InputAction}\}, \text{OutputMode})$$

$$\text{InputAction} := (\text{name}, \text{kind}, \text{dims}, \text{frame})$$

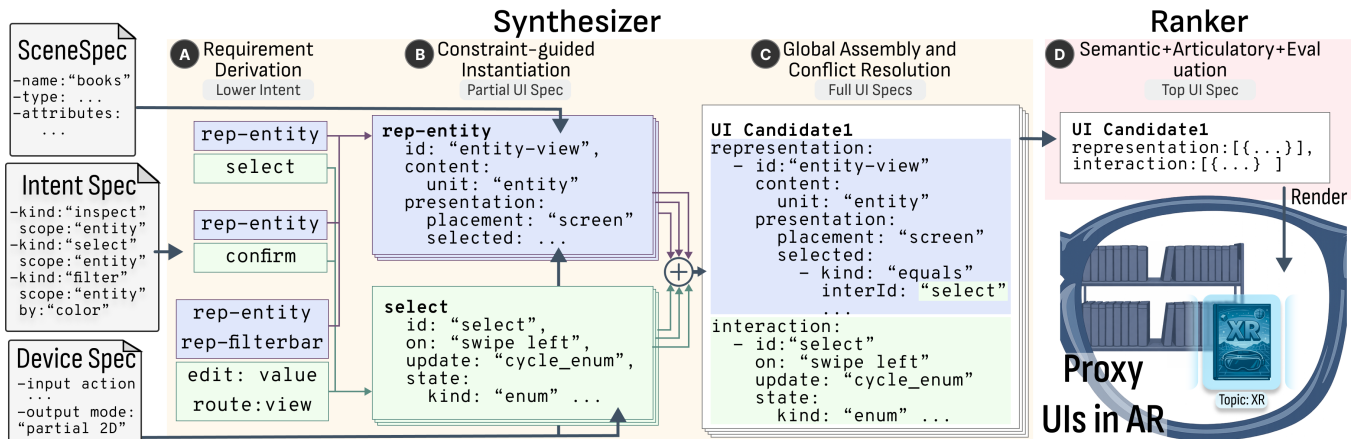
$$\text{OutputMode} \in \{\text{world}, \text{screen}, \text{partial screen}, \text{non-visual}\}$$


Figure 2: Pipeline overview with a running bookshelf example. From scene, intent, and device specifications, the synthesizer derives requirements, instantiates locally valid representation (“rep”) and interaction components under scene and capability constraints, assembles them into full UI candidates, and the ranker selects the most direct candidate using semantic, articulatory, and evaluation distance.

InputAction abstracts an available input by its type, dimensionality (i.e., *3D*, *2D*, and *Symbolic*), and reference frame (i.e., *world*, *screen*, and *none*), while OutputMode describes how content can be presented. The choice of fields and values is informed by prior taxonomies of AR interfaces [16] and interactions [30]. The design intentionally omits hardware-specific details and retains only the properties that constrain synthesis: how expressive input is, how it is spatially grounded, and where output can appear. This allows the synthesizer to reason over heterogeneous devices at the level relevant to interface generation rather than device hardware.

4.1.4 Interface. An InterfaceSpec is where the three input specifications come together. It determines how a scene is transformed into proxy-based content and presented on a given device, and how the user can act on it through the device’s input resources:

InterfaceSpec := ({Representation}, {Interaction})

Representation := (id, content, presentation, active?)

Interaction := (id, on, update, state, active?)

Inspired by Vega-Lite [51], we adopt a declarative design for the user interfaces: A Representation separates content, which specifies *what to show*, from presentation, which specifies *how to show it*; an Interaction specifies how device inputs update interface state over time. The content may expose one or more entities or groups formed by an attribute. The presentation may arrange these proxies according to their content structure or interaction requirements and place them in world space, screen space, or omit visual placement. Together, these choices support alternatives such as world-anchored entity proxies, screen-space grouped views, and symbolic lists operated through discrete input.

We separate representation from interaction so the synthesizer can reason independently about what is shown and how it can be manipulated, while making compatibility constraints explicit. This declarative structure supports checking whether a proposed interface is realizable on the target device.

4.2 Intent-Driven Synthesis

Rather than mapping inputs to a single predetermined output, the synthesizer solves a constrained synthesis problem: the same scene, intent, and device may admit multiple valid proxy interfaces. Our key observation is that the three specifications play different roles in this process. Scene determines what real-world structure is available to expose or transform, IntentSpec determines what semantic structure and operations the interface must support, and CapabilitySpec determines which realizations are feasible on the target device. This decomposition lets us formulate synthesis as intent-driven compilation under scene and device constraints.

Requirement derivation from intent. The synthesizer begins by translating each IntentSpec into a small set of canonical requirements, such as deriving an entity-level representation together with operations like selection or confirmation (Figure 2a). This step separates user goals from interface realization.

Capability-constrained instantiation. The synthesizer then instantiates candidate representations and interactions that satisfy these requirements while remaining compatible with both the scene

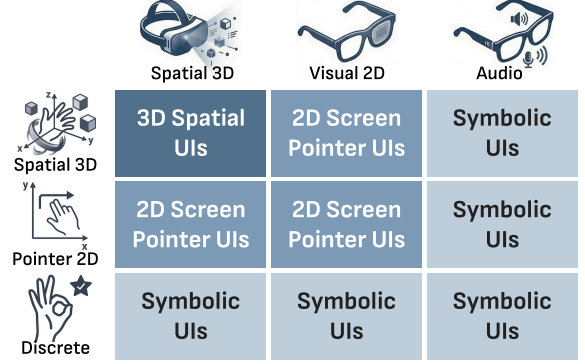


Figure 3: Interface realizability depends jointly on input and output capability. Different combinations of input and output modality admit different interface families.

and the device. Scene structure constrains what content transformations are admissible: grouping requires meaningful attributes, hierarchy operations require hierarchical structure, and proxy content must correspond to entities that actually exist. Device capability constrains how that content can be realized and manipulated. As illustrated in Figure 3, interface form depends jointly on input and output capability: richer spatial input/output combinations support 3D spatial UIs, whereas more limited combinations favor screen pointer-based or symbolic UIs. In Figure 2b, this stage produces locally valid partial UI specifications.

Global assembly and graceful degradation. Locally valid components may still conflict when combined into a complete interface. Multiple intents may compete for the same input actions, induce redundant representations, or require coordinated alternative views. We therefore formulate assembly as a bounded search over locally valid candidates, selecting combinations that maximize intent coverage while respecting device constraints and avoiding conflicts. As shown in Figure 2c, the result of this stage is a set of full UI candidates rather than a single interface. This final stage also enables graceful degradation: when all requested intents cannot be realized simultaneously, the synthesizer returns interfaces that satisfy the largest feasible subset rather than failing outright.

The output is a set of candidate InterfaceSpecs that satisfy the user’s intents for the current scene and device. Further implementation details are provided in an open-source repository <https://github.com/sys3-lab/GenerativeProxy/>.

4.3 Ranking Candidate Interfaces

Synthesis may produce multiple valid InterfaceSpecs for the same scene, device, and intents. We therefore rank candidates (Figure 2d) by operationalizing direct manipulation as three distances: *semantic execution distance*, *articulatory execution distance*, and *evaluation distance*. Intuitively, preferred interfaces expose the right task structure, make the required actions easy to perform, and make their effects easy to interpret. For each candidate, we aggregate these distances across intents using priority weights, so interfaces that better support higher-priority intents are ranked more highly.

Semantic execution distance. Semantic execution distance measures whether an interface exposes the representation needed for

an intent at the appropriate semantic level. We compute it by comparing the intent with the synthesized Representation content. Interfaces that make the required structure explicit receive lower cost; those that require the user to work through a less suitable representation receive higher cost. For example, an intent to organize objects by category is better supported by a category-grouped proxy view than by a flat list of entities.

Articulatory execution distance. Articulatory execution distance measures how difficult it is to carry out an intent with a candidate interface. We compute this distance using two components: a rule-based interaction cost and an LLM-based ergonomic score. The rule-based component captures factors directly observable from the specification, including the number of steps, required view or mode switches, input precision, and penalties for missing operations. The LLM component captures residual ergonomic properties that are difficult to formalize symbolically, such as whether an action mapping feels natural on the intent and whether the overall control scheme is coherent and discoverable. This decomposition keeps the ranking interpretable while allowing articulatory quality to reflect both measurable burden and higher-level ergonomic plausibility.

Evaluation distance. Evaluation distance measures how directly an interface makes action outcomes legible. We favor feedback embedded in, or tightly coupled to, the affected scene entities; for example, highlighting a selected object is better than showing a separate abstract state indicator. We compute this term using a rule-based tiered scheme derived from the device capability matrix (Figure 3): in-scene feedback receives the lowest cost, while screen pointer-based or symbolic feedback receives progressively higher cost as it becomes more displaced from the relevant scene object. Thus, evaluation distance reflects the spatial and semantic gap between feedback and its referent.

Finally, we apply a small shortlist-only LLM coherence penalty to top-ranked candidates to capture whole-interface inconsistencies that are difficult to detect from individual interactions in isolation. Overall, the ranker combines deterministic task- and capability-based scoring with targeted LLM judgment for ergonomic plausibility, allowing it to prefer interfaces that are not only valid, but also more direct in practice.

5 Implementation

We implement Generative Proxy as a proof-of-concept technique probe with three components: a synthesis-and-ranking backend, a runtime pipeline for acquiring scene specifications, and device-specific runtimes that render synthesized interfaces and bind them to available input/output capabilities. In the current implementation, IntentSpec and CapabilitySpec are specified in advance by developers, whereas SceneSpec is acquired at runtime when the user invokes proxy generation. This separation gives developers control over the supported intent vocabulary and target devices while allowing the generated interface to adapt to the current scene.

Synthesis and ranking backend. We implement the synthesizer and ranker in TypeScript as a Node.js web application. Both operate over JSON-based specifications: IntentSpec and CapabilitySpec are provided in advance, while Scene is acquired at runtime. The backend combines deterministic rules with LLM calls. Rule-based

logic handles specification validation, candidate synthesis, compatibility checking, and parts of distance scoring, while LLM calls are reserved for judgments that are difficult to formalize symbolically, such as ergonomic plausibility and semantic enrichment of scene entities. The backend outputs synthesized InterfaceSpecs as JSON for device-specific runtimes.

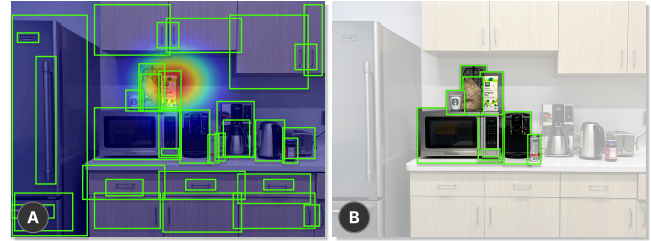


Figure 4: Gaze estimation narrows the scene for synthesis by filtering detected objects. (A) Gaze heatmap. (B) Objects retained after filtering.

Runtime scene specification acquisition. Because proxy synthesis depends on the current environment, the scene specification must be constructed online. Following Reality Proxy [38], we begin with object detection over the live egocentric view when the user triggers proxy UI generation. Whereas the original pipeline filters detections using eye gaze, many lightweight AR devices do not provide direct eye tracking or reliable head-gaze input. We therefore replace direct gaze sensing with a state-of-the-art gaze prediction model [29] that estimates gaze region from a short sequence of head-motion images. We intersect the predicted gaze region with detected bounding boxes to obtain the subset of objects used to construct the Scene specification (Figure 4), retaining only boxes intersecting regions whose predicted gaze score exceeds 0.75. We further use LLMs (GPT 5.5) to enrich detections with semantic attributes and recursive object detection to recover containment hierarchy.



Figure 5: Different AR display and input devices used in our prototypes. From left to right, the first row shows Apple Vision Pro, Meta Quest 3, Meta Ray-Ban Display, and Meta Project Aria glasses; the second row shows Apple Watch, iPhone, and Meta Neural Band.

Runtimes for different devices. To realize synthesized interfaces on heterogeneous hardware, we implement runtimes for Apple Vision Pro, Meta Quest 3, and Ray-Ban Meta glasses (Figure 5), covering both spatial and screen-space proxy interfaces. Across these platforms, we support three input families: spatial gesture input, touchscreen input, and discrete gesture input. Spatial gesture input is supported on Vision Pro and Quest 3, while touchscreen

input is provided through an iPhone companion app connected via a cloud server and discrete gesture input comes from Meta’s wristband. This prototype demonstrates feasibility of the capability class, although it does not yet match the fluidity of a dedicated commercial wristband. Overall, our implementation is intended to demonstrate feasibility of the synthesis abstraction rather than optimize for any single device.

System Response Time. Because only the SceneSpec is acquired when proxy generation is invoked, the end-to-end response time therefore comprises three components: (a) runtime scene-spec acquisition from the egocentric view, (b) candidate synthesis, and (c) candidate ranking. We benchmarked the pipeline using the two scenes from our user study (Sec. 8.2). The average end-to-end latency was below 6 s for both scenes: 4.22 ± 0.31 s for Bookshelf and 5.55 ± 0.23 s for Makerspace. The deterministic candidate-synthesis step (b) was lightweight, with latency dominated by the ML/LLM calls in steps (a) and (c). These results suggest that the current prototype is sufficiently responsive for on-demand proxy generation.

6 Application Scenarios

We use three scenarios to probe two aspects of our contribution. First, they illustrate that proxy-based UIs can support rich interaction with real-world objects across a broader range of AR devices than direct target interaction alone. Second, they show that Generative Proxy can generate different proxy UIs for the same scene and intent under different device capability constraints. The novelty illustrated by these scenarios lies not in individual operations such as filtering, semantic grouping, or hierarchical navigation, but in synthesizing proxy UIs that enable these operations across heterogeneous AR capability profiles. Accordingly, we present the scenarios as design probes rather than polished end-user applications. Together, they span dense object disambiguation on lightweight glasses, semantic regrouping for robot-mediated interaction, and multiscale navigation in building-scale environments.



Figure 6: Scenario 1 — retrieving bookshelf item information using generated symbolic Proxy UIs, corresponding to discrete input and visual 2D output. A-D show the interface on a partial 2D display, while E shows the interface on a full 2D display, where leader lines connect interface elements to their corresponding physical objects.

6.1 Scenario 1: Everyday Information Retrieval

This scenario examines dense everyday scenes in which direct target selection is difficult. Suppose a user wants to retrieve the price of a yellow-covered book. On lightweight glasses, users may lack ray casting or eye tracking to precisely select one item from many visually similar books on a shelf. The challenge is therefore not only information retrieval, but also referential disambiguation.

Proxy-based UIs address this by replacing direct target selection with a generated representation over the visible objects. Instead of selecting a specific book in place, the user interacts with a compact proxy list that can be traversed using left–right gestures (Figure 6A). The user can then open an attribute filter such as *Type*, *Color*, or *Topic* (Figure 6B). Selecting *Yellow* reduces the candidate set to matching items (Figure 6C), after which the user can inspect the remaining items and retrieve the price of the target book (Figure 6D). Under the same scene and intent, Generative Proxy can generate proxy UIs with spatial links for full display AR glasses (Figure 6E).

This scenario illustrates the proxy-UI claim that dense object selection can be redirected into a more manageable proxy space on devices with limited display area and low-precision input. At the system level, it shows that Generative Proxy can generate a compact, filterable proxy view tailored to those capability constraints.

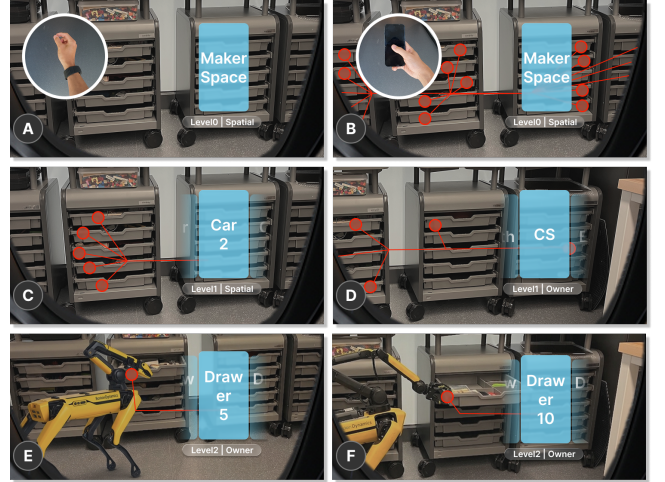


Figure 7: Scenario 2 — human–robot interaction in a makerspace using generated Proxy UIs. Panel A shows symbolic UIs corresponding to discrete input and partial 2D output, while Panel B–F show 2D screen-pointer UIs corresponding to 2D pointer input and full 2D output.

6.2 Scenario 2: Human-Robot Collaboration

This scenario examines a different challenge: the physical organization of the environment may not match the semantic structure needed for the task. In a maker space, a user wants to collaborate with a robot to retrieve materials from storage drawers based on semantic groupings such as ownership. However, the drawers are physically distributed across carts and shelves, making direct reference through speech or pointing awkward and difficult.

Proxy-based UIs address this by generating alternative structures over the same scene, intent, and across devices (Figure 7A–B).

Rather than preserving only the original spatial hierarchy, the system can regroup drawers by task-relevant attributes such as ownership. This allows the user to operate on the semantic abstraction needed for the task, while the robot maps the selected proxy back to the corresponding physical drawers. For example, in Figure 7C, the default proxy reflects the original spatial organization of the carts. The user then switches to an ownership-based organization (Figure 7D), which surfaces groups such as CS and highlights the associated drawers. After selecting the CS group, the user issues a command to request specific materials, and the robot retrieves items from the corresponding drawers (Figure 7E–F).

This scenario illustrates that proxy-based UIs support more than easier selection under sensing or ergonomic constraints: they can also expose task-relevant semantic structure that is absent from the physical arrangement itself. For Generative Proxy, it highlights the ability to favor embedded feedback that preserves the link between proxies and physical referents when device capabilities permit.

6.3 Scenario 3: Building Facility Management

This scenario examines large-scale environments in which direct interaction with individual physical referents is impractical. Building management tasks often require users to reason across multiple levels of structure, such as areas, sides, and rooms, and to reorganize that structure by semantic state rather than by physical layout alone. A direct one-object-at-a-time interaction model does not scale well to this setting.

Proxy-based UIs address this by generating hierarchical proxies that support both multiscale navigation and semantic re-leveling. For example, on partial-display glasses, the user navigates a compact location hierarchy from the lab space to a particular side of the floor (Figure 8A–B). Because proxies are generated from scene, intent, and capability specifications, the same underlying scene and task can be realized differently on different devices. Under a more immersive capability profile, Generative Proxy generates a spatially embedded proxy view of the same building structure (Figure 8C), making room-level relationships visible at a glance rather than through sequential hierarchical traversal.

More broadly, this scenario shows that proxy-based UIs can scale from local object collections to building-scale scenes, while Generative Proxy adapts their form to the target device’s capabilities.

7 Internal Ranking Sanity Check

We conducted an internal blinded ranking analysis as a sanity check on the ranking pipeline. Our goal was to assess whether the pipeline generally surfaced sensible candidates and to identify where its judgments broke down, rather than to provide definitive validation or broad comparative evaluation. No single neutral baseline applies across tasks and device profiles, since hand-designed and rule-based alternatives embed their own design choices. We therefore report success rates and representative failure cases, leaving controlled, task-specific comparisons to future work.

We constructed a factorized benchmark of 50 cases spanning 2 scene types (*art gallery* and *widget hierarchy*), 5 device profiles (*I:discrete gesture*, *O:partial visual 2D*; *I:pointer 2D*, *O:full visual 2D*; *I:discrete gesture*, *O:spatial 3D*; *I:spatial 3D*, *O:full visual 2D*;

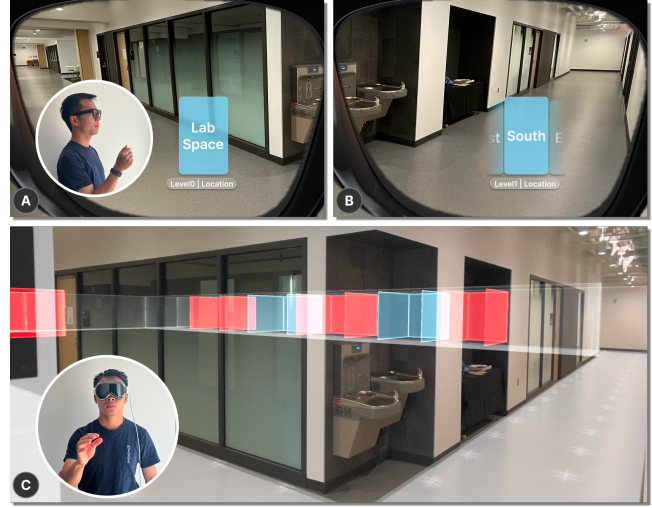


Figure 8: Scenario 3 – indoor building facility management using generated Proxy UIs. Panels A–B show symbolic UIs corresponding to discrete gesture input and partial 2D output, while panel C shows spatial 3D UIs corresponding to spatial gesture input and spatial 3D output.

I:spatial 3D, *O:spatial 3D*), and 5 intent bundles (*inspect+select*, *inspect+select+single filter*, *inspect+select+double filter*, *inspect+select+organize+relevel*, and *all intents*). For each case, we synthesized all viable interface candidates, ranked them with our method, and sampled three candidates from different parts of the ranked list: top, middle, and lower. These candidates were anonymized and presented together with the corresponding scene description, device capability profile, and ordered task intents. Two authors who were not involved in building the synthesizer or ranker independently ranked the three candidates, represented as UI specifications, from best to worst according to how well they supported the intended task. We evaluated specifications rather than deployed AR interfaces because reviewing all $50 \times 3 = 150$ candidates in-headset would have been fatiguing and impractical. We then compared the manual rankings with the system ranking using top-pick and full-order agreement; a representative packet appears in Appendix A.

Results. For Coder 1, the system matched the top-ranked candidate in 98% of cases and the full ordering in 66%. For Coder 2, top-pick agreement was 76% and full-order agreement was 62%. Inter-rater agreement between the coders was $\kappa = 0.64$, indicating substantial consistency for this diagnostic analysis.

We examined cases in which the system’s top choice did not match the coders’ top choice and found three recurring causes. First, **preference divergence in interaction decomposition** (7/14): the system preferred interfaces that merged *select* (or *focus*) and *inspect* into a single action to reduce articulatory distance, whereas coders often preferred separating these steps. Second, **ambiguity from gesture reuse across representations** (3/14): the same gesture was reused across different representations or states; although valid in the specification, coders judged these mappings as potentially ambiguous, especially without access to the realized interface. Third, **near-ties and residual evaluator disagreement**

(4/14): the remaining mismatches appeared to reflect near-ties, subjective preference, or occasional evaluator oversight rather than a systematic ranking failure.

Overall, this analysis suggests that the current pipeline is effective at surfacing strong top candidates while also revealing actionable opportunities for improvement, particularly in modeling preferences around interaction decomposition and penalizing potentially ambiguous gesture reuse.

8 Expert Evaluation

We position Generative Proxy as a proof-of-concept and technology probe, aimed at assessing whether generated proxy UIs are plausible and useful under varied AR capability conditions. To assess the perceived usefulness and limitations of the generated interfaces, we conducted an expert evaluation similar to prior technology-probe and expert-review studies in HCI [19, 38]. To compare with existing speech-based reference interaction [35] on lightweight AR glasses, we also conducted a supplemental study.

8.1 Participants and Apparatus

The study procedure and materials received IRB approval. We recruited 14 experienced XR developers and researchers from a local university’s VR clubs and research groups. Two participants joined pilot sessions used to refine the protocol and were excluded from the reported ratings, leaving 12 participants for analysis. 6 of them had more than five years of AR/VR experience, 3 had more than two years, and 11 rated their headset proficiency above 8 on a 10-point scale. Each participant received \$30 for a 90-minute session.

The study was conducted in a controlled indoor environment using a Meta Quest 3 running a Unity application. Although we implemented multiple device-specific runtimes, we used Quest 3 as a common study platform to simulate the target input/output conditions. We chose this because our Apple Watch IMU prototype was sufficient to demonstrate feasibility but did not provide interaction fluidity comparable to a dedicated wrist-worn input device; using it directly would therefore have confounded the discrete gesture condition with prototype limitations rather than the interaction concept itself. To simulate AR glasses conditions, we constrained the Quest 3 interface to the same effective capability profile, including screen-space rendering, matched display size, and the corresponding input mappings. Thus, all study conditions shared the same synthesis backend and application logic, while differing only in the realized capability condition. All sessions were recorded in audio and video, including both first- and third-person perspectives, for post-study analysis.



Figure 9: (A) Participant completing the bookshelf item-selection task. (B) Participant completing the makerspace material-finding task. (C) Supplemental study scene featuring a near-field tabletop and a far-field bookshelf.

8.2 Study Procedure

Each session lasted approximately 90 minutes and consisted of training, task trials, and post-task feedback. During training, participants were introduced to the goals and the proxy-UI interaction concepts, then completed a guided tutorial using a simplified example until they felt comfortable with the interaction techniques.

Participants then completed two tasks: (1) *Bookshelf Item Selection*, adapted from the bookshelf scenario (subsection 6.1), in which they filtered shelf items and selected items under a budget constraint; and (2) *Maker Space Material Finding*, adapted from the maker-space scenario (subsection 6.2), in which they identified drawers belonging to a designated cart and drawers associated with a target group. Each participant performed both tasks under two I/O conditions: (a) partial 2D display (simulated Meta Ray Ban Display) with micro-gesture input and (b) full 2D display with mobile touchscreen input. This yielded four trials per participant (2 tasks $\times$ 2 I/O conditions), with order counterbalanced using a Latin square.

After each trial, participants completed a 7-point Likert questionnaire assessing the UI. They also ranked the UI with another two candidates sampled in advance from the middle and bottom of the system ranking for the corresponding scene and device-capability condition. We then conducted a semi-structured interview to collect qualitative feedback on the usefulness, limitations, and potential application scenarios of Generative Proxy.

8.3 Expert Feedback

Figure 10 shows that experts rated the generated proxy UIs positively across both tasks and both device conditions. To understand the reasons behind these ratings, we analyzed the interview data using reflexive thematic analysis [2]. Two authors collaboratively coded the interview transcripts and grouped the resulting codes into three themes. Unless otherwise noted, the ratings discussed below are averaged across the four study cases.

8.3.1 Overall usability and utility of proxy-based UIs on constrained AR devices. Experts generally found proxy-based UIs easy to understand (6.5/7), easy to use (6.25/7), controllable (6.17/7), and effective across tasks (6.40/7). These results suggest that proxy interaction remained usable under the simulated lightweight-glasses capability constraints. Participants particularly valued compact proxy representations when limited display area and low-precision input made spatial target selection cumbersome (P3, P9). Rather than pointing precisely at distant or crowded objects, users could navigate, filter, and select within a compact interface. Attribute filters narrowed visually similar items in the bookshelf task (P4, P11), while hierarchical grouping exposed task-relevant structure in the maker-space task (P5). Overall, participants regarded proxy UIs as a practical way to support real-world object interaction when device capabilities do not support precise spatial targeting.

8.3.2 Feedback on generated proxy UIs across device configurations. Experts also commented on whether the generated proxy UIs appropriately matched different device conditions. Across conditions, ratings for *easy to tell success* remained consistently positive (6.52/7), suggesting that the generated UIs generally made action outcomes legible. Ratings for *feel natural* (5.92/7) and *match what I think* (6.04/7) were also positive overall, but varied more across

conditions, indicating that these dimensions were more sensitive to the fit between proxy form and device capability.

Table 1: Agreement between system and expert rankings. Top Pick indicates agreement on the best candidate; Full indicates exact agreement on the full three-way ranking.

Scenario	Condition	Top Pick	Full
Bookshelf	Partial + Gesture	83.33%	58.33%
Bookshelf	Full + Touch	91.67%	58.33%
Maker Space	Partial + Gesture	91.67%	58.33%
Maker Space	Full + Touch	100%	100%

We also compared the system’s ranking of candidate UIs with expert judgments. As shown in Table 1, the system’s top-ranked candidate matched the experts’ top choice in 83.33% to 100% of cases, while full ranking agreement ranged from 58.33% to 100%. Agreement was strongest in the *Full + Touch* makerspace condition, where the system exactly matched expert preferences. Overall, these results suggest that the ranker captures expert preference reasonably well, especially when identifying the best candidate, whereas differences between middle- and lower-ranked alternatives were less distinct.

Qualitative feedback helps explain this pattern. When device capabilities allowed stronger visual linkage to the physical environment, participants preferred proxies with linked highlights, dots, and lines, which made referents easier to identify and action outcomes easier to interpret (P4, P5, P8). P4 noted, “*Linking the proxy to real-world objects really helps convey the association*”. In more constrained partial-display conditions, participants (e.g., P7) still found the generated proxies useful, but relied more on compact organization, filtering, and hierarchical navigation when direct visual correspondence was limited. This helps explain why evaluation-related ratings remained strong across conditions, while naturalness and semantic fit were more mixed in the constrained cases.

Experts also noted that different device conditions called for different proxy forms. Compact, filterable views were seen as appropriate for *Gesture + Partial* interaction (P2, P3, P9, P12), where display space and input bandwidth were limited, whereas *Touch + Full* conditions benefited from richer spatial correspondence and more persistent links to scene objects (P2, P6, P8). At the same time, participants (P6, P7, P11) appreciated that core operations such as traversing, filtering, and selecting remained conceptually stable across configurations. Together, feedback supports the core premise of Generative Proxy: generated proxy UIs should adapt to device capability while preserving a consistent interaction vocabulary.

8.3.3 Design opportunities and limitations. Despite the generally positive feedback, experts also identified several opportunities for improvement. First, while hierarchy and filtering were useful, participants (e.g., P12) noted that proxy structures could become harder to read in constrained partial-display conditions. In particular, the tree structure was not always immediately legible when only a small portion of the interface was visible at once. Commented by P2, “*Visible tree structure would help map my mental model with the physical model*.” This suggests a need for stronger visual support for hierarchy, such as clearer level indicators, progressive disclosure, or lightweight overview cues. Second, participants emphasized that feedback should remain as tightly coupled to the physical scene as possible. When device capabilities supported embedded visual correspondence, experts preferred in-scene or scene-linked feedback because it made action outcomes easier to interpret. P6 highlighted that “*I definitely liked the lines going to the drawers. It helped me really understand what the hierarchy is showing me*.” When such feedback was absent or reduced, the interaction remained usable, but felt more abstract and required more cognitive effort (P11). This reinforces the importance of evaluation legibility in generated proxy UIs.

8.4 Supplemental Study

To examine when proxy-based interaction is preferable to direct reference on lightweight AR glasses, we conducted a supplemental within-subject study comparing proxy UIs generated by Generative Proxy with direct speech-based reference. We used speech as a representative direct-reference modality that allows users to identify real-world targets without first navigating an intermediate interface [35]. Seven XR experts from the participant pool described in Sec. 8.1 evaluated both methods in two settings (Figure 9C): a sparse, near-field tabletop scene (0.5m) and a dense, far-field bookshelf scene (3m).

Participants completed information retrieval tasks by filtering objects and retrieving prices encoded in their metadata. In the speech condition, participants referred directly to target objects using speech; in the proxy condition, they navigated and selected from a generated proxy UI using the Meta Neural Band. Both conditions used the Meta Ray-Ban Display. Each session lasted approximately 30 mins, including training, tasks, and a semi-structured interview.

Six of seven participants preferred direct speech-based reference in the sparse tabletop setting, whereas six of seven preferred proxy-based interaction in the dense bookshelf setting. Participants explained that speech was efficient when objects were clearly visible and easy to name, and that it offered flexibility for open-ended

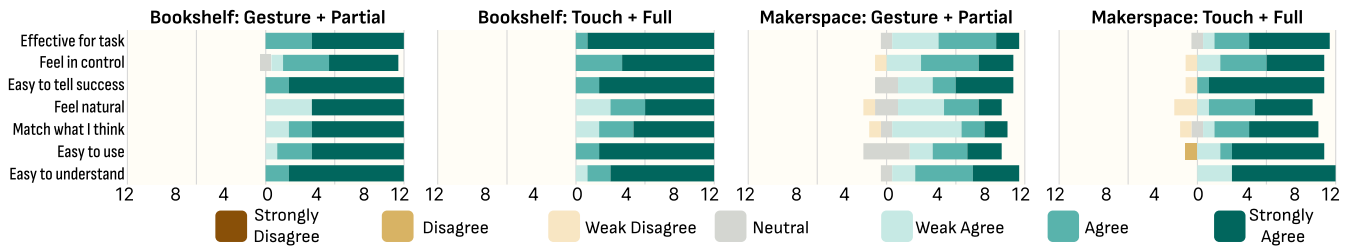


Figure 10: User feedback on UIs generated by Generative Proxy across different scene, intent, and device specifications.

requests (E2, E3). In contrast, they preferred proxy UIs when objects were crowded, difficult to distinguish, or difficult to describe precisely, because the proxy provided explicit alternatives for disambiguation (E3, E7).

Because the two settings jointly varied distance, density, and target nameability, these results do not isolate the effect of each factor independently. Instead, they provide initial evidence that proxy-based interaction complements direct speech-based reference on lightweight AR glasses: speech may be more natural when targets are readily perceivable and nameable, whereas proxies may be more useful when direct reference is ambiguous or cumbersome.

9 Discussion

9.1 Proxy UIs as a Capability-bridging Interaction Layer

Our results suggest that the key value of proxy UIs is not only that they avoid difficult direct targeting, but that they introduce a digital interaction layer that is more fluid than the physical scene itself. Proxies can be reorganized, regrouped, and re-leveled at runtime to match the user’s task and the capabilities of the current device. This makes them a capability-bridging layer: they preserve the semantics of real objects while freeing interaction from the fixed layout of the environment and the limitations of any single device.

This perspective helps explain why proxy form should vary across platforms. Richer devices can support tighter scene linkage and embedded feedback, while more constrained devices benefit from compact panels, hierarchical navigation, or symbolic interaction. More broadly, proxy-based UI should be understood not as a workaround for limited AR hardware, but as a general strategy for transforming real-world interaction into adaptive digital representations that remain grounded in the physical world.

9.2 Generating AR UIs as Miniatures of the User’s Mental Model

Although we do not formulate Generative Proxy as a generative model in the probabilistic sense, the interface is synthesized at runtime from the current scene, user intent, and device capability, and LLM-based judgments in the ranking stage. In this sense, the system does not retrieve a fixed interface template, but generates a proxy UI specific to the current situation.

More importantly, our results suggest that effective generative UI in AR is not only a matter of producing functionally correct widgets or layouts. Because the interface mediates interaction with the physical world, it must also reflect how users mentally organize that world for the task at hand. In our scenarios, the most useful proxies were those that exposed task-relevant structure already implicit in the environment, such as grouping books by topic, drawers by ownership, or spaces by hierarchical location. These proxies function as interactive miniatures of the user’s working mental model: compact representations that preserve the distinctions most relevant for action on a given device.

This suggests that generative UI in AR should be understood not simply as interface generation, but as the synthesis of interactive representations that align the user’s mental model, the physical scene, and the device’s capabilities. A key next step is therefore

richer user modeling, so that generated proxies can better reflect differences in users’ goals, prior knowledge, and preferred ways of structuring the world.

9.3 Limitations and Future Work

Our study used Meta Quest 3 as a controlled simulator for multiple capability conditions. Future work should validate the system on a broader set of native deployments. Moreover, the current latency of proxy UI generation is compatible with the intended use of Generative Proxy for user-triggered, on-demand synthesis rather than continuous per-frame generation. Future work could further reduce latency through caching and parallelization, as well as by replacing cloud-based LLM calls with faster local or task-specific models. Our expert evaluation focused on experienced XR users, most of whom were university-affiliated. Broader studies are needed to understand how proxy-based UIs generalize to less technical users and to longer-term everyday use. Proxy-based UI is better understood as an interaction paradigm than as one low-level technique, and its appropriate comparison point may vary across tasks and devices. Future work should compare generated proxies against task-specific direct interaction baselines where appropriate. Finally, our current implementation focuses on visual and gesture-based interaction and does not yet explore voice as an input modality. Voice may be especially valuable for lightweight AR devices and could complement proxy navigation in future systems.

10 Conclusions

We introduced Generative Proxy, a method for synthesizing proxy-based interfaces for interacting with real-world objects across heterogeneous AR capability conditions. Grounded in constrained synthesis and a ranking procedure based on semantic, articulatory, and evaluation distance, Generative Proxy takes structured specifications of the scene, the user’s intent, and the target device to generate proxy-based interfaces. By decoupling interaction from the physical constraints of the environment and the limitations of any single device, these interfaces enable users to interact more effectively with objects and spaces that are dense, distant, cluttered, or otherwise difficult to access through direct AR interaction. Through several application scenarios, we demonstrated the versatility of this approach across diverse scenes, intents, and AR form factors. Our internal ranking sanity check and expert evaluation further highlight the promise of generated proxy UIs as a capability-bridging interaction layer for AR, suggesting a path toward more fluid, adaptive, and expressive interaction with real-world objects across the next generation of AR glasses.

Acknowledgments

We thank the reviewers for their constructive feedback. This work is partially supported by NSF 2543054 and Google.

References

- [1] Karan Ahuja, Sujeeth Paredy, Robert Xiao, Mayank Goel, and Chris Harrison. 2019. Lightanchors: Appropriating point lights for spatially-anchored augmented reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 189–196.
- [2] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

- [3] Yining Cao, Peiling Jiang, and Haijun Xia. 2025. Generative and malleable user interfaces with generative and evolving task-driven data model. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [4] Baptiste Caramiaux, Marco Donnarumma, and Atsuo Tanaka. 2015. Understanding gesture expressivity through muscle sensing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 6 (2015), 1–26.
- [5] Ishan Chatterjee, Jiexin Ding, Anandghan Waghmare, Joseph Breda, Yuquan Deng, Bo Liu, Yuntao Wang, and Shwetak Patel. 2025. FlowRing: Integrated Microgesture and Surface Interaction Ring for Versatile XR Input. *Proceedings of the ACM on Human-Computer Interaction* 9, 5 (2025), 1–28.
- [6] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+ gesture: Expressive, precise and targeted free-space interactions. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 131–138.
- [7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [8] Yifei Cheng, Yukang Yan, Xin Yi, Yuanchun Shi, and David Lindlbauer. 2021. Semanticadapt: Optimization-based adaptation of mixed reality layouts leveraging virtual-physical semantic connections. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 282–297.
- [9] Mustafa Doga Dogan, Eric J Gonzalez, Karan Ahuja, Ruofei Du, Andrea Colaco, Johnny Lee, Mar Gonzalez-Franco, and David Kim. 2024. Augmented object intelligence with xr-objects. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.
- [10] Ruofei Du, Alex Olwal, Mathieu Le Goc, Shengzhi Wu, Danhang Tang, Yinda Zhang, Jun Zhang, David Joseph Tan, Federico Tombari, and David Kim. 2022. Opportunistic interfaces for augmented reality: Transforming everyday objects into tangible 6dof interfaces using ad hoc ui. In *CHI Extended Abstracts*. 1–4.
- [11] João Marcelo Evangelista Belo, Mathias N Lystbæk, Anna Maria Feit, Ken Pfeuffer, Peter Kán, Antti Oulasvirta, and Kaj Grønbaek. 2022. Auit—the adaptive user interfaces toolkit for designing xr applications. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [12] Jacqui Fashimpaur, Tovi Grossman, Ben Lafreniere, Naveen Sendhilnathan, Kashyap Todi, Tianyi Wang, Ting Zhang, and Tanya R Jonker. 2025. Squiggle: Multimodal Lasso Selection in the Real World. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [13] Andreas Fender, Philipp Herholz, Marc Alexa, and Jörg Müller. 2018. OptiSpace: Automated Placement of Interactive 3D Projection Mapping Content. In *Proc. of CHI*. ACM, New York, 269. doi:10.1145/3173574.3173843
- [14] Andreas Fender, David Lindlbauer, Philipp Herholz, Marc Alexa, and Jörg Müller. 2017. HeatSpace: Automatic Placement of Displays by Empirical Analysis of User Behavior. In *Proc. of UIST*. ACM, New York, 611–621. doi:10.1145/3126594.3126621
- [15] Figma, Inc. 2026. *Figma*. <https://www.figma.com/> Collaborative interface design tool.
- [16] James D Foley. 1996. *Computer graphics: principles and practice*. Vol. 12110. Addison-Wesley Professional.
- [17] Mar Gonzalez-Franco and Andrea Colaco. 2024. Guidelines for productivity in virtual reality. *Interactions* 31, 3 (2024), 46–53.
- [18] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. 2016. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE transactions on visualization and computer graphics* 23, 6 (2016), 1706–1724.
- [19] Han L Han, Miguel A Renom, Wendy E Mackay, and Michel Beaudouin-Lafon. 2020. Textlets: Supporting constraints and consistency in text documents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [20] Fengming He, Xiyun Hu, Xun Qian, Zhengzhe Zhu, and Karthik Ramani. 2024. AdaptUI: Adaptation of Geometric-Feature-Based Tangible User Interfaces in Augmented Reality. *Proceedings of the ACM on Human-Computer Interaction* 8, ISS (2024), 44–69.
- [21] Anders Henrysson, Mark Billinghurst, and Mark Ollila. 2005. Face to face collaborative AR on mobile phones. In *Fourth IEEE and ACM international symposium on mixed and augmented reality (ismar'05)*. IEEE, 80–89.
- [22] Anuruddha Hettiarachchi and Daniel Wigdor. 2016. Annexing reality: Enabling opportunistic use of everyday objects as tangible proxies in augmented reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1957–1967.
- [23] Gaoping Huang, Xun Qian, Tianyi Wang, Fagun Patel, Maitreya Sreeram, Yuanzhi Cao, Karthik Ramani, and Alexander J. Quinn. 2021. AdapTutAR: An Adaptive Tutoring System for Machine Tasks in Augmented Reality. In *Proc. of CHI*. ACM, New York, 417:1–417:15. doi:10.1145/3411764.3445283
- [24] Edwin L Hutchins, James D Hollan, and Donald A Norman. 1986. Direct manipulation interfaces. In *User centered system design*. CRC Press, 87–124.
- [25] Yue Jiang, Ruofei Du, Christof Lutteroth, and Wolfgang Stuerzlinger. 2019. ORC layout: Adaptive GUI layout with OR-constraints. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*. 1–12.
- [26] Tai-Wei Kan, Chin-Hung Teng, and Wen-Shou Chou. 2009. Applying QR code in augmented reality applications. In *Proceedings of the 8th international conference on virtual reality continuum and its applications in industry*. 253–257.
- [27] Hirokazu Kato and Mark Billinghurst. 1999. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Proceedings 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99)*. IEEE, 85–94.
- [28] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A Lee, and Mark Billinghurst. 2018. Pinpointing: Precise head-and-eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [29] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. 2024. In the eye of transformer: Global-local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision* 132, 3 (2024), 854–871.
- [30] Joseph J LaViola Jr, Ernst Kruijff, Ryan P McMahan, Doug Bowman, and Ivan P Poupyrev. 2017. *3D user interfaces: theory and practice*. Addison-Wesley Professional.
- [31] Geonsun Lee, Min Xia, Nels Numan, Xun Qian, David Li, Yanhe Chen, Achin Kulshrestha, Ishan Chatterjee, Yinda Zhang, Dinesh Manocha, et al. 2025. Sensible agent: A framework for unobtrusive interaction with proactive ar agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–22.
- [32] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [33] Lik Hang Lee, Tristan Braud, Farshid Hassani Bijarbooneh, and Pan Hui. 2020. UbiPoint: towards non-intrusive mid-air interaction for hardware constrained smart glasses. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 190–201.
- [34] Lik-Hang Lee, Tristan Braud, and Pan Hui. 2023. Embodied interaction on constrained interfaces for augmented reality. In *Springer Handbook of Augmented Reality*. Springer, 239–271.
- [35] Lik-Hang Lee and Pan Hui. 2018. Interaction methods for smart glasses: A survey. *IEEE access* 6 (2018), 28712–28732.
- [36] Alan Leung, Ruijia Cheng, Jason Wu, Jeffrey Nichols, and Titus Barik. 2025. SQUIRE: Interactive UI Authoring via Slot QQuery Intermediate REpresentations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–17.
- [37] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 147–160.
- [38] Xiaolan Liu, Difan Jia, Xianhao Carton Liu, Mar Gonzalez-Franco, and Chen Zhu-Tian. 2025. Reality proxy: fluid interactions with real-world objects in MR via abstract representations. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [39] Xianhao Carton Liu, Difan Jia, Tongyu Nie, Evan Suma Rosenberg, Victoria Interrante, and Chen Zhu-Tian. 2026. Can AR Embedded Visualizations Foster Appropriate Reliance on AI in Spatial Decision-Making? A Comparative Study of AR X-Ray vs. 2D Minimap. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [40] Yuwen Lu, Alan Leung, Amanda Swearngin, Jeffrey Nichols, and Titus Barik. 2025. Misty: UI prototyping through interactive conceptual blending. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [41] Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems* 77, 12 (1994), 1321–1329.
- [42] Bryan Min, Allen Chen, Yining Cao, and Haijun Xia. 2025. Malleable overview-detail interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.
- [43] Darius Miniotas, Oleg Spakov, Ivan Tugoy, and I Scott MacKenzie. 2006. Speech-augmented eye gaze interaction with small closely spaced targets. In *Proceedings of the 2006 symposium on Eye tracking research & applications*. 67–72.
- [44] Kyzyl Monteiro, Ritik Vatsal, Neil Chulpongsoatorn, Aman Parnami, and Ryo Suzuki. 2023. Teachable reality: Prototyping tangible augmented reality with everyday objects by leveraging interactive machine teaching. In *Proc. CHI*. 1–15.
- [45] Richard Nguyen, Charles Gouin-Vallerand, and Maryam Amiri. 2023. Hand interaction designs in mixed and augmented reality head mounted display: a scoping review and classification. *Frontiers in Virtual Reality* 4 (2023), 1171230.
- [46] Jeffrey Nichols, Brad A Myers, and Kevin Litwack. 2004. Improving automatic interface generation with smart templates. In *Proceedings of the 9th international conference on Intelligent user interfaces*. 286–288.
- [47] Benjamin Nuernberger, Eyal Ofek, Hrvoje Benko, and Andrew D Wilson. 2016. Snapreality: Aligning augmented reality to the real world. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1233–1244.
- [48] Ken Pfeuffer, Hans Gellersen, and Mar Gonzalez-Franco. 2024. Design principles and challenges for gaze+ pinch interaction in xr. *IEEE Computer Graphics and Applications* 44, 03 (2024), 74–81.

- [49] Xun Qian, Fengming He, Xiyun Hu, Tianyi Wang, Ananya Ipsita, and Karthik Ramani. 2022. Scalar: Authoring semantically adaptive augmented reality experiences in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [50] Yutong Ren, Arnav Reddy, and Michael Nebeling. 2026. PeriphAR: Fast and Accurate Real-World Object Selection with Peripheral Augmented Reality Displays. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [51] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2016. Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 341–350.
- [52] Ryo Suzuki, Mar Gonzalez-Franco, Misha Sra, and David Lindlbauer. 2025. Everyday AR through AI-in-the-Loop. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–5.
- [53] Ryo Suzuki, Rubaiat Habib Kazi, Li-yi Wei, Stephen DiVerdi, Wilnot Li, and Daniel Leithinger. 2020. Realitysketch: Embedding responsive graphics and visualizations in AR through dynamic sketching. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 166–181.
- [54] Amanda Swearngin, Chenglong Wang, Alannah Oleson, James Fogarty, and Amy J Ko. 2020. Scout: Rapid exploration of interface layout alternatives through high-level design constraints. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
- [55] Pedro Szekely, Ping Luo, and Robert Neches. 1993. Beyond interface builders: Model-based interface tools. In *Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems*. 383–390.
- [56] Tomu Tahara, Takashi Seno, Gaku Narita, and Tomoya Ishikawa. 2020. Retargetable AR: Context-aware Augmented Reality in Indoor Scenes based on 3D Scene Graph. In *Proc. of ISMAR*. IEEE Computer Society, Los Alamitos, 249–255. doi:10.1109/ISMAR-Adjunct51615.2020.00072
- [57] Wai Tong, Chen Zhu-Tian, Meng Xia, Leo Yu-Ho Lo, Linping Yuan, Benjamin Bach, and Huamin Qu. 2022. Exploring interactions with printed data visualizations in augmented reality. *IEEE transactions on visualization and computer graphics* 29, 1 (2022), 418–428.
- [58] Ching-Yi Tsai, Nicole Tacconi, Andrew D Wilson, and Parastoo Abtahi. 2026. Uncertain Pointer: Situated Feedforward Visualizations for Ambiguity-Aware AR Target Selection. *arXiv preprint arXiv:2602.13433* (2026).
- [59] Priyan Vaithilingam, Elena L Glassman, Jeevana Priya Inala, and Chenglong Wang. 2024. Dynavis: Dynamically synthesized ui widgets for visualization editing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [60] Priyan Vaithilingam and Philip J Guo. 2019. Bespoke: Interactively synthesizing custom GUIs from command-line applications by demonstration. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*. 563–576.
- [61] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Ke Huo, Yuanzhi Cao, and Karthik Ramani. 2020. CAPtURAR: An augmented reality tool for authoring human-involved context-aware applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 328–341.
- [62] Zeyu Wang, Ruotong Yu, Xiangyang Wang, Jiexin Ding, Jiankai Tang, Jun Fang, Zhe He, Zhuojun Li, Tobias Röddiger, Weiye Xu, et al. 2025. Computing with smart rings: A systematic literature review. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (2025), 1–54.
- [63] Matt Whitlock, Ethan Harnner, Jed R Brubaker, Shaun Kane, and Danielle Albers Szafir. 2018. Interacting with distant objects in augmented reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 41–48.
- [64] Jason Wu, Eldon Schoop, Alan Leung, Titus Barik, Jeffrey P Bigham, and Jeffrey Nichols. 2024. Uicoder: Finetuning large language models to generate user interface code through automated feedback. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 7511–7525.
- [65] Brad Vander Zanden and Brad A Myers. 1990. Automatic, look-and-feel independent dialog creation for graphical user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 27–34.
- [66] Chen Zhu-Tian, Daniele Chiappalupi, Tica Lin, Yalong Yang, Johanna Beyer, and Hanspeter Pfister. 2023. RL-LABEL: A Deep Reinforcement Learning Approach Intended for AR Label Placement in Dynamic Scenarios. *IEEE transactions on visualization and computer graphics* 30, 1 (2023), 1347–1357.
- [67] Chen Zhu-Tian, Yijia Su, Yifang Wang, Qianwen Wang, Huamin Qu, and Yingcai Wu. 2020. MARVisT: Authoring Glyph-Based Visualization in Mobile Augmented Reality. *IEEE Trans. Vis. Comput. Graph.* 26, 8 (2020), 2645–2658. doi:10.1109/TVCG.2019.2892415
- [68] Chen Zhu-Tian, Wai Tong, Qianwen Wang, Benjamin Bach, and Huamin Qu. 2020. Augmenting Static Visualizations with PapARVis Designer. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

APPENDIX

A Representative Ranking Sanity-Check Case

Figure 11 presents one representative packet from the 50-case internal ranking sanity check. We selected Case 25, which uses a flat art-gallery scene with spatial 3D input and full visual 2D screen output. Its ordered intents cover inspection, selection, and primary and secondary filtering.

The evaluation interface preserves the view presented to the coders: three anonymized candidate proxy UIs (A–C) are shown side by side with their representation choices and gesture-to-intent mappings. The participant was asked to rank these three options based on their personal preference. This example illustrates how valid candidates for the same scene, device profile, and intent specification can differ in interaction assignment and state-update operations.

The complete scene, device capability, intent, and InterfaceSpec JSON for this case, together with the remaining ranking packets, is provided in `ranking_sanity_check_supplement.zip` in the supplementary materials.

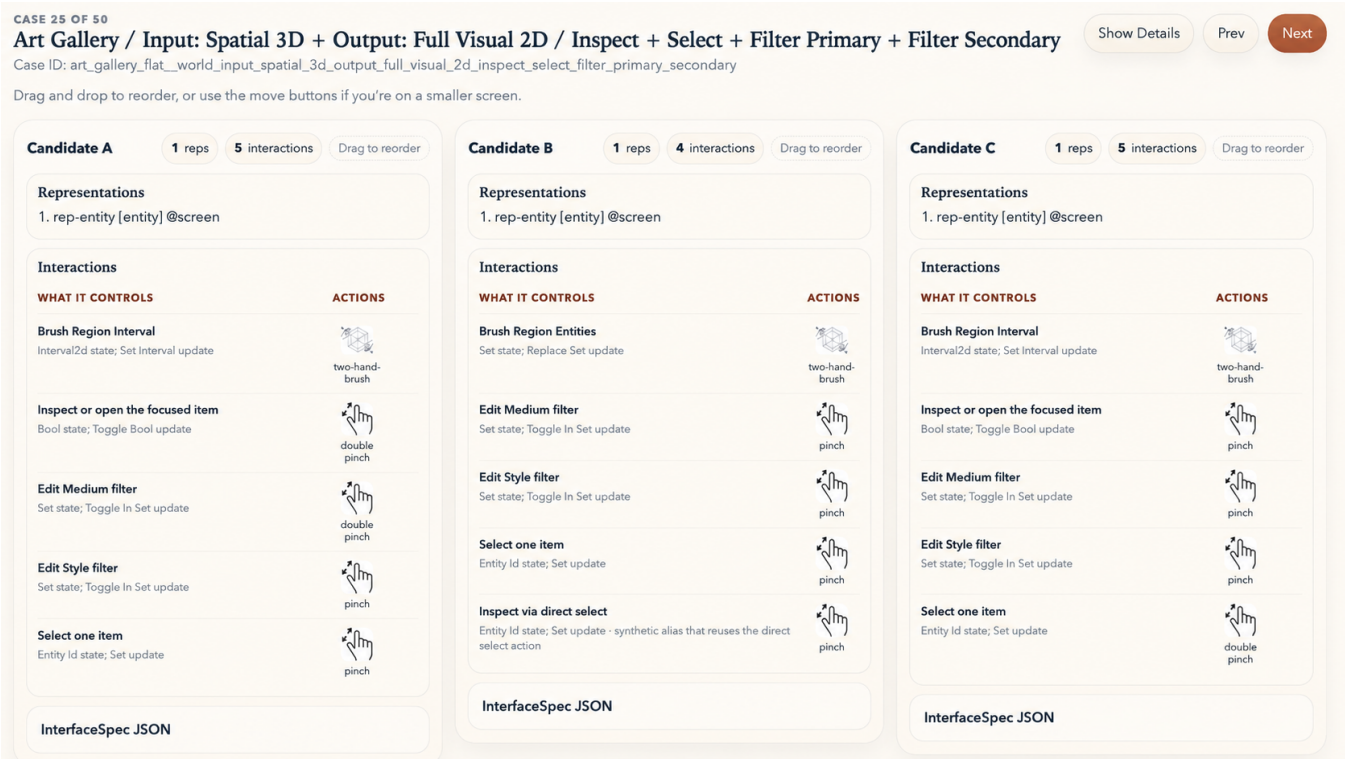


Figure 11: Representative candidate set for Case 25 in the internal ranking sanity check. The art-gallery case uses spatial 3D input and full visual 2D screen output for inspection, single- and multi-selection, and primary and secondary filtering. The evaluation interface presents three anonymized candidate proxy UIs side by side.