Augmented Object Intelligence: Making the Analog World Interactable with XR-Objects

Mustafa Doga Dogan
MIT CSAIL, USA
Google, Switzerland

Eric J. Gonzalez
Google, USA

Karan Ahuja
Northwestern University, USA
Google, USA

Ruofei Du
Google, USA

Andrea Colaço
Google, USA

Johnny Lee
Google, USA

Mar Gonzalez-Franco*
Google, USA

David Kim*
Google, Switzerland

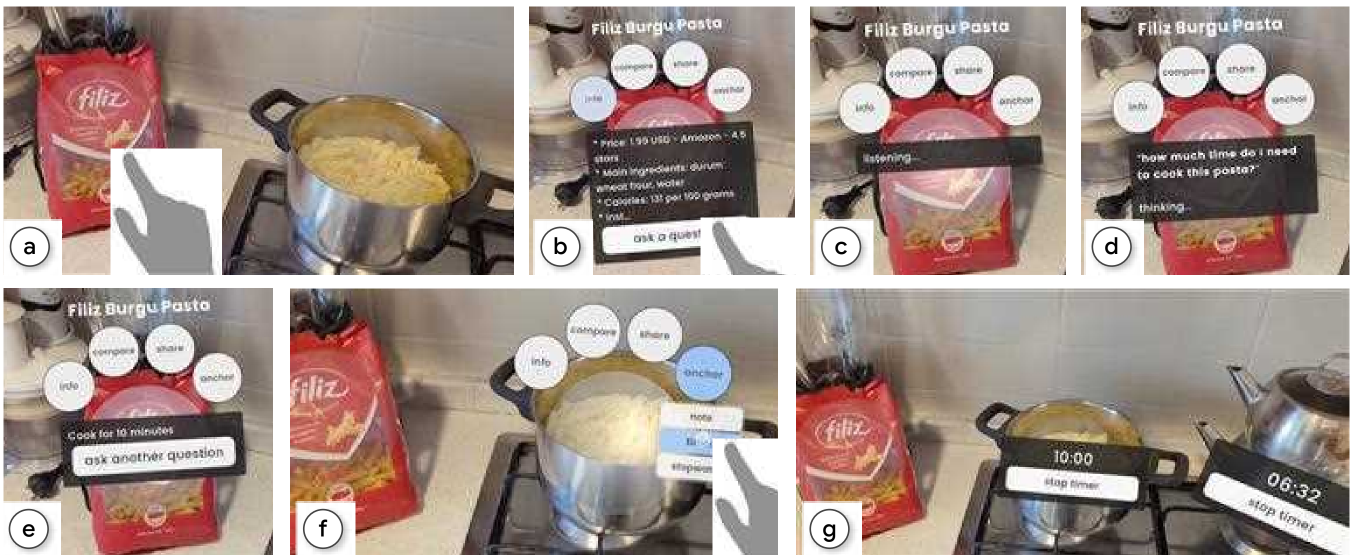

Figure 1: XR-Objects allows users to (a) select and interact with real-world objects in AR as if they were digital objects. (b) Automatically generated object-based AR context menus allow objects to provide information about themselves, such nutritional facts. (c, d, e) For example, a user asks a question about cooking time, and (f,g) uses the result to set a spatial timer widget anchored to the relevant pot in 3D space.

ABSTRACT

Seamless integration of physical objects as interactive digital entities remains a challenge for spatial computing. This paper introduces Augmented Object Intelligence (AOI), a novel XR interaction paradigm designed to blur the lines between digital and physical by equipping real-world objects with the ability to interact as if they were digital, where every object has the potential to serve as a portal to vast digital functionalities. Our approach utilizes object segmentation and classification, combined with the power of Multimodal Large Language Models (MLLMs), to facilitate these interactions. We implement the AOI concept in the form of XR-Objects, an open-source prototype system that provides a platform for users to engage with their physical environment in rich and contextually relevant ways. This system enables analog objects to not only convey information but also to initiate digital actions, such as querying for details or executing tasks. Our contributions are threefold: (1) we define the AOI concept and detail its advantages over traditional AI assistants, (2) detail the XR-Objects system's open-source design and implementation, and (3) show its versatility through a variety of use cases and a user study.

*Both authors share co-senior authorship.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Mixed / augmented reality**; • **Computing methodologies** → **Intelligent agents**.

KEYWORDS

mixed reality; extended reality; augmented reality; augmented objects; spatial computing; user interfaces; context menus

1 INTRODUCTION

Modern Extended Reality (XR) platforms come with a plethora of sensors, cameras, and advanced computer vision techniques to

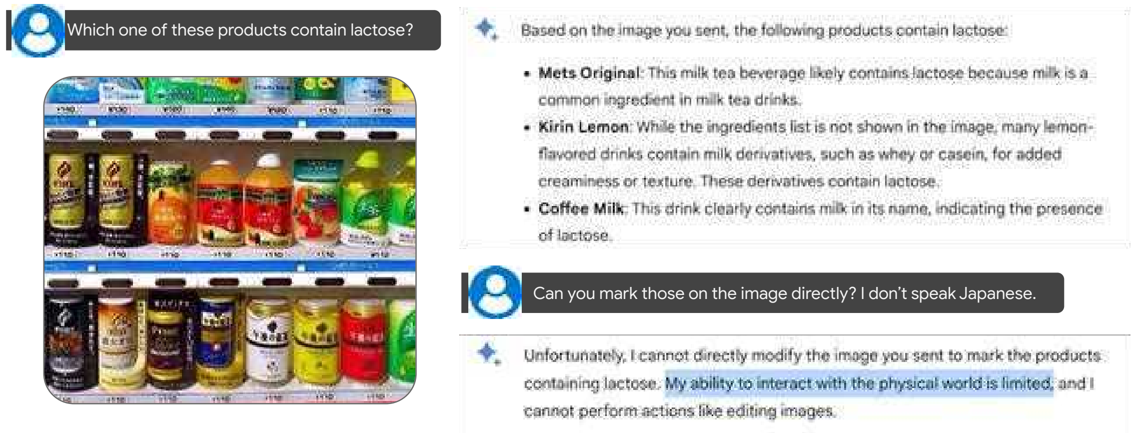

Figure 2: Example of an interaction with a conventional multimodal AI assistant. While the model clearly has the capacity to produce reasonable scene understanding when an image and a prompt is provided as input, it fails in providing an anchored output that ties back to the original multimodal prompt.

seamlessly blend virtual content with the physical world through color passthrough and scene understanding. However, despite these technological steps, the integration of real objects into the XR environment remains somewhat superficial, treating the physical world largely as a mere backdrop rather than an interactive component. In contrast, projects like *RealityCheck* [28] or *Remixed Reality* [43] show a future where digital and physical worlds could be closely intertwined together. Similarly, advancements in artificial intelligence (AI) are laying the groundwork for such a future, with breakthroughs in real-time unsupervised segmentation [55] combined with in-painting [61] or generative AI content generation [34].

The wide availability of machine learning (ML) and computer vision technologies has also led to features that enhance digital interaction with the physical world at our fingertips. Tools like image-based search in Google Lens[1] and utility-focused Augmented Reality (AR) features in smartphones, such as text copy & paste and real-time translation, are becoming increasingly common. Together, these tools form building blocks that could bring us closer to a future in which a total understanding of the world and its objects can be applied to our everyday interactions in XR.

In this paper, we explore a novel interaction paradigm we term *Augmented Object Intelligence (AOI)* that would allow any real object identified by the XR system to reveal digital data associated with the **analog object** and perform context-appropriate **digital actions** in a meaningful way. Our system XR-Objects embodies this idea and aims to demonstrate and investigate "semantic equality" between real and virtual objects.

Imagine a scenario as familiar as right-clicking a digital file to open its context menu, but applied to physical objects within XR — such as right-clicking on potatoes or pasta in a pot to start a cooking timer set to the correct duration, or filtering for gluten-free products on a grocery shelf through an XR interface (see Figure 1).

The leap towards physical awareness in XR-Objects represents an advancement over traditional AR, which often relies on manual input or the use of physical tracking markers. We leverage developments in spatial understanding via technologies such as Simultaneous Localization and Mapping (SLAM), available in *ARCore*[2] and *ARKit*[3], and machine learning models for object segmentation and classification (*COCO* [40] via *MediaPipe* [44]). These technologies enable us to implement AR interactions with semantic depth. We also integrate a Multimodal Large Language Model (MLLM) into our system, which further enhances our ability to automate the recognition of objects and their specific semantic information within XR spaces.

Our contributions are threefold:

- We introduce the concept of Augmented Object Intelligence (AOI), a paradigm shift towards seamless integration of real and virtual content in XR.
- We detail the open-sourced design and implementation of XR-Objects, our prototypical system that exemplifies AOI, alongside an exploration of diverse use cases to demonstrate its potential.
- We provide a comparison between standard prompt-based LLM interfaces and our AOI approach for contextual information retrieval and object-centric interaction.

We demonstrate our prototype that integrates these AR and AI components in a seamless way, which is implemented for smartphones to provide access to a broad audience, as commercial headsets do not yet give programmatic access to the user's camera stream. By open-sourcing our system, we aim to foster further innovation in the field, ultimately bringing us closer to a future where the physical and digital realms interact seamlessly, embodying the concept of Programmable Reality [23] for everyday users.

2 RELATED WORK

This section provides an overview of previous work in the realms of user interface (UI) design principles, blended realiry interactions,

[1]Google Lens: https://lens.google

[2]Google ARCore: https://developers.google.com/ar

[3]Apple ARKit: https://developer.apple.com/augmented-reality/arkit

and advancements in AR technology [51]. These areas form the foundation upon which our research builds, which aims to enhance the interaction between users and the physical world through XR-Objects.

2.1 Fundamentals of UI

The challenge of bridging the cognitive gap between human users and computational systems has been a central theme in HCI. Traditional approaches have employed various layers of representations to mediate this interaction, manifesting most recognizably in the UIs of devices ranging from PCs and smartphones to IoT devices and automotive systems [56]. Despite these advancements, the resurgence of command-line interface (CLIs) in contemporary AI interactions, as seen in the usage of prompts with large language models (LLMs) [52], suggests a potential oversimplification of user interaction paradigms. This regression shows the necessity of reevaluating our approach to UI design in the age of AI and spatial computing [53], where the intricacy of human intentions and the computational interpretation thereof demand a more nuanced form of representation. Figure 2 demonstrates that the multimodal AI assistant clearly has the capacity to produce reasonable scene understanding when an image and a prompt is provided as input, but it fails in providing an anchored output that ties back to the original multimodal prompt.

Significantly, the role of UIs extends beyond mere facilitation of interaction, which shapes the user's ability to navigate, understand, and command the computational system. Effective UIs support essential cognitive functions such as memory, discovery, and articulation [5], thus becoming a key factor in the widespread adoption and utility of AI and spatial computing technologies.

In addressing these challenges, our work revisits the utility of context menus — a familiar paradigm in desktop computing [2, 4, 65] — and explores their potential in fostering familiar interactions with physical objects within XR environments. Previous demonstrations [38] showed the potential of such ubiquitous context menus through simulated digital notes via projection. In this work, we show a fully functional prototype using AR.

2.2 Blended Reality

The exploration of blended reality systems is a rapidly developing field within HCI, marked by recent research that integrates real and virtual environments [28, 42, 43]. While a comprehensive literature review [3] exceeds the scope of this section, we draw attention to key themes that emerged within the broader XR literature.

Manipulating Perception in XR: Gonzalez-Franco and Lanier [26] explored how perceptual manipulations can be effectively modeled within VR environments to enrich user experience. Bonnail et al. explore how XR can leverage human memory limitations to influence user perception and behavior [6]. Concurrently, *causality-preserving asynchronous reality* enables users to interact with events in a causally accurate manner despite temporal delays [22]. Tseng et al. highlighted the risks associated with such manipulations [58] and proposed mitigations against their malicious use [57].

Interacting with Real-World Affordances: As XR platforms become increasingly accessible, enhancing physical-aware interactions can significantly elevate user experiences. *DepthLab* [21] exemplifies advancements in real-time 3D interactions with the real world using depth maps. *InteractionAdapt* [11] optimizes VR workspaces for efficient interaction across diverse physical environments, demonstrating the flexibility and adaptability of VR systems to support varied user activities.

Balancing Immersion and External Awareness: Critical to blended-reality systems is managing the balance between immersion in the virtual world and awareness of the real world. Studies such as those by Kudo et al. emphasize the importance of smoothly transitioning users and devices between realities, while also maintaining bystander awareness [36]. Further guidelines on this balance are discussed in works by Gonzalez-Franco et al. for harmonizing user experience across dimensions [24, 25].

2.3 Interacting with Physical Objects in XR

Despite the considerable advances in AI's capability to generate and understand complex content, our physical world remains predominantly analog, with only a fraction of daily activities and tools being enhanced by digital technology [49]. This analog nature of human experiences, from basic needs fulfillment to complex task execution, presents a significant challenge in integrating digital intelligence in a manner that feels both natural and easy to grasp.

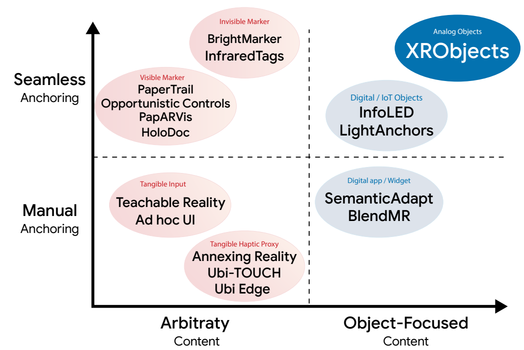

Figure 3: The landscape of physical object interactions in XR classified across two dimensions: anchoring and content.

Several research efforts have explored ways to bridge this gap by leveraging XR technologies to enable interaction with real-world objects. Figure 3 shows the landscape of physical object interactions in XR classified across two dimensions: anchoring (manual vs. seamless) and content (arbitrary vs. object-focused).

Some approaches rely on pre-registration of objects or manual setups to achieve tangible input [20, 45, 66] or tangible haptic proxy [29, 31, 33]. In contrast to these manual processes, other works utilized markers to more automatically execute the object detection and AR content anchoring. For instance, researchers embedded visible [9, 30, 39, 48] or invisible markers [17–19] to documents and 3D objects for AR purposes. However, these methods often impose limitations on the types of objects that can be interacted with, requiring specific fabrication or preparation.

Our exploration acknowledges the complexity of translating digital interactions to physical objects and aims to bridge this divide by enhancing any physical object with digital functionality and contextual interaction capabilities.

Once the objects have been identified and localized by the AR platform, an important dimension is what *content* will be shown at the identified locations [7]. Previous work has focused on optimizing the presentation of digital content (e.g., existing app windows or widgets) on or around objects [10, 27, 41], while others have investigated the use of tangible interaction with physical objects as a means to control such digital content [45, 54, 66]. *EditAR* [13] suggested capturing users and their interactions with objects nearby to create digital twins for later consumption in XR. *ProcessAR* [12] captures instruction demonstrations related to domain-specific objects by experts, so they can be viewed by novices in situ.

InfoLED [63], *LightAnchors* [1], and *Reality Editor* [32] took this one step further to show truly *object-specific content* in AR, however, this was specifically for electronic objects, such as their showing their battery level or device status in AR. Researchers further suggested augmenting visualizations with dynamic AR content [9, 14, 15, 62], however, such dynamic content was limited to printed documents or displays.

Our work builds upon these prior efforts by proposing a system that enables **object-specific content** and interactions for a much wider range of objects, regardless of their physical capabilities or pre-configured markers. We leverage advancements in spatial understanding via techniques like SLAM [16], available in *ARCore* and *ARKit*, and machine learning models for object segmentation and classification to achieve this. This allows us to implement AR interactions with semantic depth, enabling contextually relevant information and actions for any object in the user's environment.

3 XR-OBJECTS IMPLEMENTATION

XR-Objects leverages developments in spatial understanding via tools such as SLAM, available in *ARCore*[4] and *ARKit*[5], and machine learning models for object segmentation and classification (*COCO* [40] via *MediaPipe* [44]), which enables us to implement AR interactions with semantic depth. We also integrate a Multimodal Large Language Model (MLLM) into our system, which further enhances our ability to automate the recognition of objects and their specific semantic information within XR spaces.

Platform. Given the current constraints of AR headsets, particularly their limited developer access to real-time camera streams, we consciously targeted smartphones for our mobile prototype development. We aim to enable anyone to try out our open-source project on their phone. Modern smartphones use similar types of ARM-based mobile chipsets as AR headsets, yielding comparable performance for real-time computer vision tasks, while providing unrestricted access to their high-resolution cameras. This enables our application to identify objects in the user's environment and overlay digital information directly onto the physical world through the phone's display thanks to *ARCore* and *ARKit*.

Multimodal Interaction. At the heart of XR-Objects is a multimodal large language model (MLLM) [64] as well as a speech recognizer, which facilitate a rich interaction layer between the user and the objects. This model not only recognizes objects but also fetches and provides contextual information and actions relevant to the selected object. By integrating voice and visual inputs, our system offers a seamless and familiar interface for users to engage with their surroundings in novel ways.

[4]Google ARCore: https://developers.google.com/ar
[5]Apple ARKit: https://developer.apple.com/augmented-reality/arkit

3.1 Design Considerations

In developing the AOI paradigm for XR environments, we considered a range of design choices to enhance user interaction and system performance. These considerations are grounded in our review of related work and guided by our goal to seamlessly integrate digital functionalities with physical objects. Here, we explain our rationale behind key design decisions, contrasting them with alternative approaches and situating them within the broader discourse of HCI and AR.

3.1.1 Object-Centric vs. App-Centric Interaction. Traditional AR interactions often follow an app-centric model, where users must first open a specific application to access digital functionalities. In this model, users have to navigate through the app's interface to select categories or objects of interest, and in some cases, even upload pictures for analysis. Examples of app-centric interactions include standard *ChatGPT*-style interfaces, where users input a query and an image, and *Google Lens*, which requires users to open the app and manually select the objects they wish to interact with.

In contrast, our system prioritizes an object-centric approach, where interactions are directly anchored to objects within the user's environment. This means that users can immediately access digital functionalities by selecting an object, without the need to navigate through an app or input additional information. By leveraging advanced computer vision and spatial understanding techniques, our AOI framework enables users to seamlessly engage with the physical world as if it were a digital interface.

The object-centric approach offers several advantages over app-centric models. Firstly, it provides a more natural interaction flow, as users can directly engage with objects in their surroundings without the cognitive burden of switching between the physical world and a digital app. Secondly, it minimizes the operational steps required to access digital functionalities, streamlining the user experience and reducing friction. Finally, by anchoring interactions directly to objects, our AOI framework fosters a more immersive and seamless XR experience.

3.1.2 World-Space UI vs. Screen-Space UI. The choice between implementing a world-space UI versus a screen-space UI was informed by our aim to maintain spatial consistency and enhance user engagement with the XR environment. A screen-space UI, fixed relative to the user's viewpoint, could potentially obfuscate the immersive experience by detaching digital interactions from their physical context. Conversely, our adoption of a world-space UI, where digital elements are anchored to physical objects (akin to "billboards" in 3D graphics, i.e., user-facing 2D planes in a 3D space), ensures that interactions remain contextually grounded within the user's real-world environment. We hope to minimize cognitive load by preserving spatial orientation and also leverage the natural human capability to navigate and interact with 3D spaces.

3.1.3 Signaling Identified XR-Objects. To mitigate visual clutter, a common issue in densely populated AR environments, we introduce the use of semi-transparent spheres, or "bubbles," as minimalist

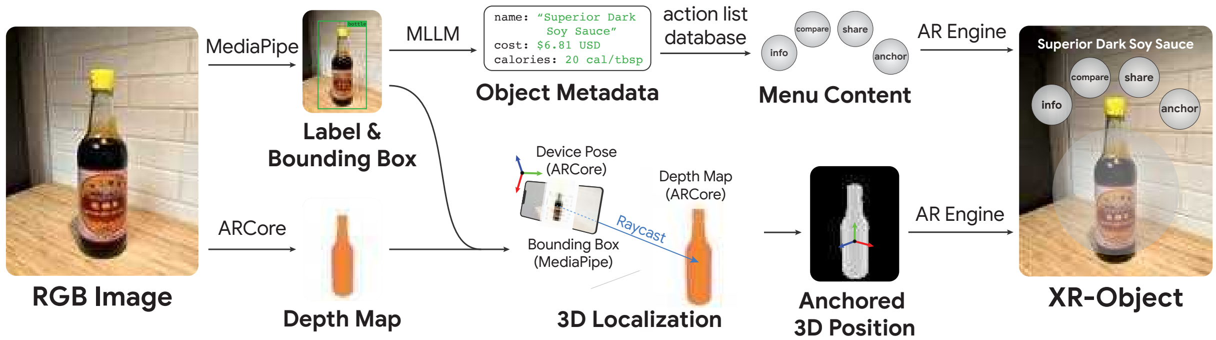

Figure 4: The XR-Objects processing pipeline combines MediaPipe and ARCore for object detection and spatial tracking, respectively, integrates an MLLM for object-specific metadata retrieval and interaction, and renders UI content in 3D space.

indicators of interactable objects. This design choice is based on the principle of minimalism and unobtrusiveness, ensuring that users are not overwhelmed by excessive digital information overlaying their physical surroundings. Bubbles serve as subtle prompts that an object is interactive to balance informational availability with spatial aesthetics.

3.1.4 Fixed Number of Top-Level Categories and Actions. The decision to implement a fixed number of top-level categories and actions within the system's UI was driven by considerations of usability and cognitive efficiency. Limiting the choice set helps mitigate decision fatigue and simplifies the interaction process, making it easier for users to navigate the system's functionalities. This design philosophy aligns with the Hick-Hyman Law [60], which states that increasing the number of choices proportionally increases decision time. By streamlining the number of options available, we are also able to adopt a radial menu with constant reach distance [47, 50] instead of dropdown lists, and we facilitate quicker user decision-making and enhance the overall user experience.

In summary, we aim to deliver a seamless, efficient, and immersive XR experience by opting for an object-centric interaction model, employing a world-space UI, utilizing visual bubbles for indicating interactability, and limiting the complexity of user choices.

3.2 Categories of Actions

Our system facilitates fluid interactions with a single or multiple objects and enables users to take various digital actions, such as querying real-time information, asking questions, sharing the objects with contacts, or adding spatial notes. Inspired by sub-menus in traditional context menus on desktop computing, we categorized our seven implemented actions into four categories, which we list below.

(1) **Information**: provide an overview; ask a question
(2) **Compare**: ask to compare multiple objects within the view
(3) **Share**: send object to a contact; add to shopping list
(4) **Anchor**: notes; timer; countdown

In the above list, the first two categories (**Information** and **Compare**) represent traditional *Visual Question Answering (VQA)* tasks, while the other two (**Share** and **Anchor**) represent traditional *widget* tasks. We open-source our code on *GitHub*[6] and anticipate that the list of integrated actions will be extended in the future by the XR community.

3.3 System Architecture

The implementation of XR-Objects involves a series of steps to augment real-world objects with functional context menus as illustrated in Figure 4. These steps can be categorized as (1) detecting the objects, (2) localizing and anchoring onto the object, (3) coupling each object with an MLLM for metadata retrieval, and (4) on user input, executing the actions and displaying the output. We use *Unity* and its *AR Foundation*[7] to bring the necessary components for these steps together to build our system. Below, we detail the components and processes that constitute this framework.

3.3.1 Object Detection and Classification. The foundation of XR-Objects is its robust object detection module, which leverages the capabilities of the *Google MediaPipe* library [44]. This module employs a convolutional neural network (CNN) optimized for mobile devices, providing *on-device* and real-time classification of objects within the user's camera feed. The system detects objects by providing a *class label* (e.g., "bottle", "monitor", "plant") and generating *2D bounding boxes*, which serve as preliminary spatial anchors for subsequent AR content. The current CNN model is based on the *COCO* dataset [40] which provides 80 labels.

To prioritize user privacy and data efficiency, XR-Objects further processes only those object regions that are of relevance to the user's current interaction context. For instance, even though *MediaPipe* inherently also identifies people in the scene, a region classified as a "person" by the on-device model is not sent to the MLLM-based cloud query system to preserve the privacy of users in the surroundings. Similarly, other classes of objects that are relevant or irrelevant to the user can be customized depending on the AR application (e.g., a plant species search AR app could only run queries for the "plant" class).

3.3.2 3D Localization and Anchoring. With the object identified, XR-Objects proceeds to generate AR menus. These menus are

[6]XR-Objects open-source project: https://github.com.anonymized
[7]Unity AR Foundation: https://unity.com/unity/features/arfoundation

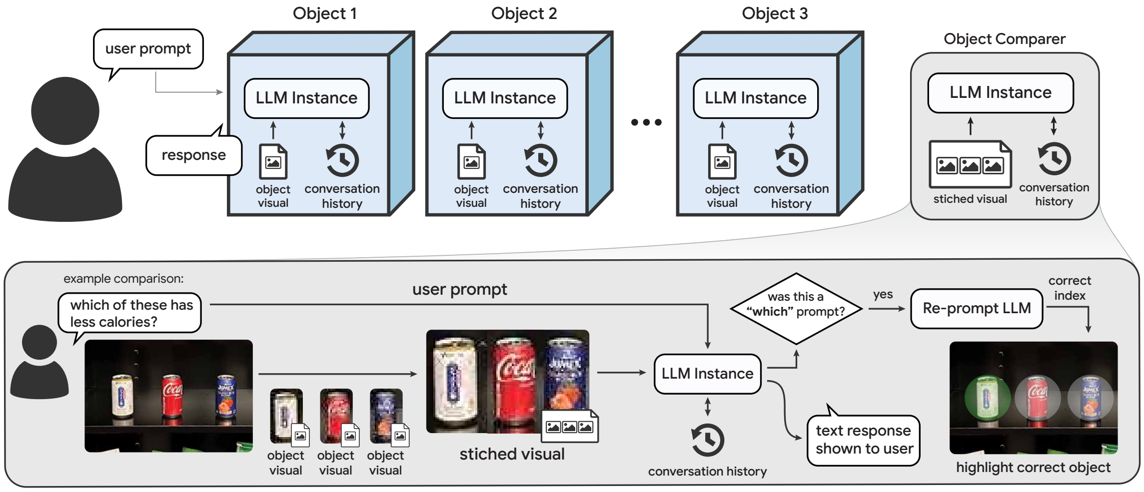

Figure 5: XR-Objects instantiates a dedicated MLLM instance for identified object in the scene. Object comparisons are executed by stitching together the relevant objects in the scene before passing the query to an MLLM instance.

spatially anchored to the objects using a combination of the initial 2D bounding boxes and depth information of the scene. We us raycasting to translate the 2D object locations into precise 3D coordinates.

In our system, we used the depth map on the phone [21] generated through depth from motion [59] by *ARCore*. Because the location returned by the object detector from the previous step is in 2D screen space, we raycast from this point toward the depth map to "hit" the object and find the corresponding 3D object location in world space, as shown in Figure 4.

At the computed 3D location, we instantiate our object proxy template, which was developed as a *prefab* (i.e., fully configured *GameObjects* saved in the AR project for reuse) in *Unity*. The AR engine ensures that the object proxy stays anchored even when the user changes their view angle.

The object proxy contains the object's context menu, however, before the user selects the object, it shows up as only a small semi-transparent sphere on the object, which signals to the user that the object has been recognized by the system. Only when the user taps this sphere, the full context menu is shown, otherwise, the menu remains hidden to avoid visual overload.

Our algorithm also includes extra steps to ensure the object proxies do not get spawned at undesired locations or get erroneously duplicated for the same object.

3.3.3 Coupling Each Object with a MLLM. We couple a MLLM with each identified object; thus, we run one LLM conversation instance per object, as shown in Figure 5. We use the cropped bounding box from the first step as the visual input to the MLLM. We also store the conversation history by referring to a conversation ID internally. This object-specific approach enables the MLLM to provide detailed information about the object that extends far beyond the capabilities of traditional object classifiers. For example, it can furnish the object with a wide array of data, including but not limited to product specifications, historical context, and user reviews. As demonstrated in Figure 4, the system is capable of recognizing an object as "Superior Dark Soy Sauce," rather than merely identifying it as a "bottle"—the generic label typically assigned by standard object detection processes in the preceding step.

For our MLLM, we use *PaLI* [8], which runs in the cloud and takes as input the captured region of interest (i.e., object's cropped image). The MLLM system is able to simultaneously query the Internet (*via Google Search*) to retrieve additional metadata about the object, e.g., prices and user ratings in the case of a product.

3.3.4 Menu Interaction. Interactions within XR-Objects are facilitated through a multimodal interface, supporting both touch and voice inputs. This flexible interaction model allows users to engage with the system in a manner that best suits their preferences and the current context. For voice interactions, the system incorporates a speech recognition engine[8], which enables the processing of natural language commands and queries. As the feedback mechanism, certain user actions, such as selecting a menu option or asking a question, are reflected in the panel overlaid on the object.

When the object is selected by the user, the actions, described in the previous section, are shown. Once an action's button is tapped, the interaction starts in a panel overlaid on the object.

Information retrieval. For the actions that retrieve real-time data (e.g., getting the answer to the user's question), we use the object's MLLM instance. For instance, when the "info" button is selected, the MLLM-returned object summary is shown. We use the following prompt to create an object summary:

> *Provide the information from the following list that makes sense for this object. Fill in the missing "…" using info from the Internet. Exclude the one that are irrelevant. Divide the relevant ones with a "*". * Price: … (give price+vendor+score/rating) * Cheaper alternatives: name - price * Main ingredients: … (top 2) * Calories: … * Allergens: … * Instructions: … (short) * Care: … (if fashion/tool/plant). Use extremely*

[8]Speech Recognizer: https://github.com/EricBatlle/UnityAndroidSpeechRecognizer

short answers and exclude answers that are 'None' or 'n/a' or 'irrelevant'. Limit to 30 words.

If the user wants to ask a more specific question, they can tap the "ask a question" button, and directly speak out a question.

Object comparisons. For the object *compare* functionality, we use a dedicated "object comparer" method, which allows us to compile multiple identified objects' information and provide the combined result as input to a dedicated MLLM instance. As shown in Figure 5, the object comparer stitches all objects' images together and provides its MLLM instance when the user asks a question $prompt_{user}$ about multiple objects using the "compare" button. The returned response is shown to the user.

If the user's prompt is a "which" question, the object comparer also executes a follow-up MLLM query under the hood to help visualize the results for this "filtering" type of user question. For this reprompting, we augment the user's prompt $prompt_{user}$ with a sub-prompt $prompt_{indexing}$ such as:

Considering that the items are ordered from left to right with the first object being index 0, tell me ONLY the correct indices, written as numbers.

Thus, the MLLM returns only the right indices, which we use to mark the relevant objects in the AR view as shown in the bottom-right screenshot in Figure 5.

4 EVALUATION

We conducted a user study comparing XR-Objects to a state-of-the-art MLLM assistant interface (Gemini app[9]), referred to as "Chatbot" from here on, for contextual information retrieval and object-centric interaction. Participants were asked to perform a number of timed VQA tasks and widget interactions in simulated grocery shopping and at-home scenarios, and provided feedback on their experience with each interface through a survey. This study was approved by Institutional Review Board at [anonymized].

4.1 Participants

We recruited 8 participants (6 male, 1 female, 1 preferred not to disclose) between the ages of 25-45. All were fluent English speakers (4 native), were regular shoppers, and all but 1 had used smartphone-based AR at least once.

4.2 Scenario & Tasks

We designed a scenario consisting of a simulated grocery shopping experience (Figure 6a) followed by an at-home experience (Figure 6b) in which users complete a set of 6 tasks using either XR-Objects or Chatbot. The task categories included: (**T1**) 2-object Search, (**T2**) N-object Search, (**T3**) Share Object, (**T4**) Get Info (Single Object), (**T5**) Set/Anchor Timer, and (**T6**) Create/Anchor Note. We created two variants of this scenario (**A** and **B**), each with the same categories but slightly varied tasks. The full set of tasks is presented in Table 1.

For XR-Objects , participants were instructed to use the **Compare** feature for **T1** and **T2**, the **Share** feature for **T3**, the **Info** feature for **T4** and the **Anchor** feature for **T5** and **T6**. For Chatbot, participants were instructed to take/upload a photo along with a query (using their preferred method of text or voice) to the chat for **T1**, **T2**, and **T4**. For the remaining tasks, participants were told to use Chatbot as if it were connected to a smartphone assistant.

	Scenario A		Scenario B
	Grocery Store:		Grocery Store:
T1.	**2-object search (Shelf 1)** *Which soy sauce has less protein?*	T1.	**2-object search (Shelf 1)** *Which oil is from Spain?*
T2.	**N-object search (Shelf 2)** *Which beans have the most fat?*	T2.	**N-object search (Shelf 3)** *Which drink has the least calories?*
T3.	**Share object** *Send a message to Mom asking if she'd like you to buy Dried Mango.*	T3.	**Share object** *Send a message to Mom asking if she'd like you to buy Dried Mango.*
	Home (same for both Scenarios):		Home (same for both Scenarios):
T4.	**Get single object info** *Find brew time for the tea*	T4.	**Get single object info** *Find brew time for the tea*
T5.	**Set/anchor a timer** *Set a timer for the tea*	T5.	**Set/anchor a timer** *Set a timer for the tea*
T6.	**Create/anchor a note** *"buy more juice"*	T6.	**Create/anchor a note** *"buy more juice"*

Table 1: Task descriptions given in the user study.

4.3 Procedure

Participants were first given a brief introduction to the study, provided informed consent, and filled out a demographics survey. The experimenter then walked through the functionality of both XR-Objects and Chatbot, and participants then completed a set of sample tasks on an object not included in the study.

Figure 6: User study setup with mock grocery store (a) and at-home (b) environments. Examples of using XR-Objects in each case are shown in (c) and (d), respectively.

Each participant completed the tasks in both scenarios **A** and **B**. The ordering of both condition (XR-Objects, Gemini) and scenario (**A**, **B**) were counterbalanced between participants to prevent ordering effects. Once the tasks were completed, participants were free to use the tool freely for up to five additional minutes however they chose. Participants then completed a post-condition survey on their qualitative experience (Appendix A.1).

[9]Google Gemini: https://play.google.com/store/apps/details?id=com.google.android.apps.bard

After a two minute break, this was repeated with the remaining condition and scenario. Upon completing the tasks with both conditions, participants completed a final survey comparing interactions with XR-Objects and Chatbot (Appendix 4.4.3). Overall, the study took approximately 45 minutes.

4.4 Measures

4.4.1 Task Completion Time. We recorded the time required to complete tasks **T1-T6** for each condition as a measure of overall system performance.

4.4.2 Post-Condition HALIE Survey. Following the completion of all tasks in a given condition (XR-Objects, Chatbot), participants completed a survey evaluating their interactions, adapted from the Human-AI Language-based Interaction Evaluation (HALIE) framework proposed by Lee et al. [37]. Participants rated their agreement with the following statements (among others) on a 5-point *Likert* scale. Due to space constraints, the complete survey is provided in Appendix A.1.

4.4.3 Post-Study Form Factor Survey. While the user study was conducted on a smartphone due to limitations of head-mounted display (HMD) camera access, our vision for XR-Objects is for it to run entirely on the HMD. Therefore, we conducted a post-study survey in which participants envisioned interacting with XR-Objects on an HMD (e.g., *Apple Vision Pro*). The survey questions (provided in Appendix A.2) were based on HALIE, but formulated as a forced-choice comparison between the Chatbot and XR-Objects interaction paradigms.

4.5 Analysis

4.5.1 Task Completion Time. We analyze completion time using traditional *t-tests*, and confirm normality via *Shapiro-Wilk Test.*

4.5.2 Distribution Skewness. Skewness (γ_1) quantifies the asymmetry of a given distribution's shape. For a normal distribution, values of $|\gamma_1| < 0.5$ indicate an approximately symmetric distribution. Values of $0.5 < |\gamma_1| < 1$ suggest moderate skewness, while $|\gamma_1| > 1$ suggests a highly skewed distribution. This statistical approach, in contrast to visual methods like histograms, is particularly useful for analyzing distributions in *Likert*-scale questionnaires [35].

4.5.3 Generalized Linear Model. To analyze the data derived from our forced-choice questionnaires, we use a Generalized Linear Model (GLM) based on the Logit Binomial distribution. Unlike regular linear models, GLMs enable regression beyond Gaussian distributions. Considering this data follows a Bernoulli distribution (i.e., datapoints are 0 or 1), our GLM is effectively a log-odds model.

4.6 Results

4.6.1 Time. On average, participants using XR-Objects required significantly less time (M=217.5 s, SD=58 s) to complete all tasks when compared to the participants using Chatbot (M=286.3, SD=71 seconds), illustrated in Figure 7 and confirmed by a paired t-test (t=-2.8, df=5, p=0.01). This translates to a roughly 31% in task completion time on average for Chatbot users compared to XR-Objects.

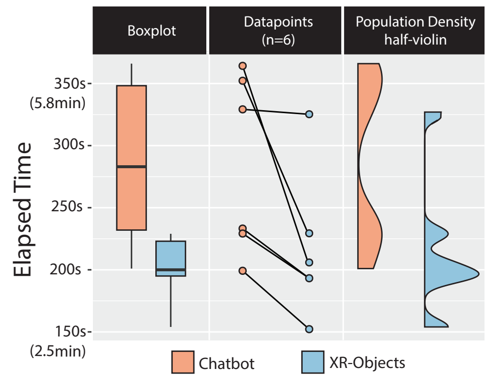

Figure 7: Comparison of task completion times between XR-Objects and Chatbot. The box plot visualizes the distribution of data, individual data points, within-subject comparisons (n=6), and the distribution for each condition.

4.6.2 HALIE Survey. We analyze the HALIE survey results using both traditional non-parametric tests for ordinal data and skewness calculations to assess the distribution of responses (Figure 8).

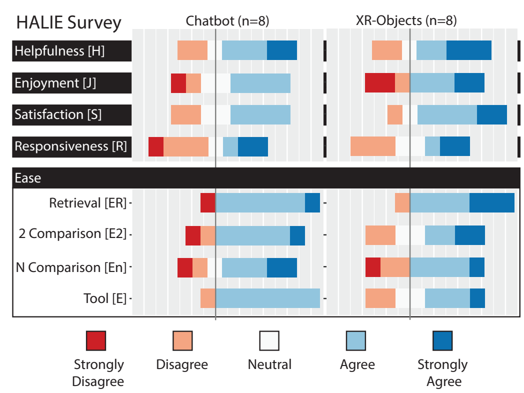

Figure 8: *Likert*-scale results of the HALIE survey.

While we find no significant differences between Chatbot and XR-Objects on any factor of the HAILE survey (*Wilcoxon Paired tests*), we find that both approaches of MLLM-enabled real-world search (either via Chatbot or XR-Objects) appear positively rated. Thus, we proceed with a skewness analysis. The most significant skewness of the questionnaire data were found on the questions regarding Ease of Use, and Satisfaction. In particular, Ease of Information Retrieval showed both conditions were highly skewed: XR-Objects ($\gamma_1 = 1.19$) Chatbot ($\gamma_1 = 1.8$), making a strong case for MLLM-enabled information retrieval in any form.

Further exploration shows that the responses for Tool Ease were highly skewed for Chatbot ($\gamma_1 = 2.25$). However, those same skews weren't sustained on the XR-Objects condition ($\gamma_1 = 0.03$). We hypothesize that this is because XR-Objects is a research prototype, while the Chatbot used was a fully released product. Nevertheless, we found a moderate skewness on the questionnaire results for the

Satisfaction metric only for our prototype XR-Objects ($\gamma_1 = 0.7$), but not for the Chatbot condition ($\gamma_1 = 0.4$).

4.6.3 Form Factor (Phone or HMD). The results of the form factor survey are summarized in Figure 9. We applied a GLM with two factors: *FormFactor* (phone, HMD) and *Question* (across 12 questionnaire levels), assuming a binomial distribution. Given that the dependent variable's responses were binary (XR-Objects or Chatbot), traditional linear models were inadequate as they are tailored to fit Gaussian distributions only. The GLM approach allowed for fitting a Bernoulli distribution and conducting appropriate tests. Our analysis revealed a significant *FormFactor* effect $F(191, 179) = 1.917, p < 7.05e - 08, \eta^2 = 3.8$.

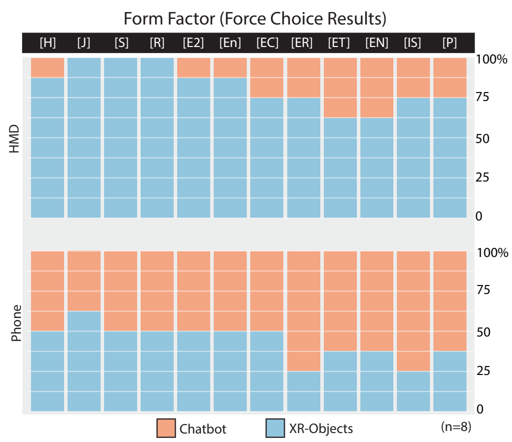

Figure 9: *Marimekko*-chart mosaic with the form factor survey results.

To further assess the model's effectiveness in predicting Form Factor, we examined the model deviance ($-2LL = 209$) and compared it against the null model's deviance, which assumed Form Factor was not a consideration ($-2LL = 253$). This comparison demonstrated that our model was more adept at accounting for the variance, where a higher deviance signifies a poorer fit. A chi-square test contrasting the two models yielded a significant difference, with $\chi^2 = 1.6 \times 10^{-5}, df = 12$.

These findings show a clear preference for XR-Objects in the context of the HMD form factor. Conversely, when using a phone, participants' preferences between the AI tools (Chatbot or XR-Objects) were split, validating our hypothesis regarding the optimal form factor for tools like XR-Objects.

5 APPLICATIONS

Through AOI, we envision XR-Objects to be useful across a variety of real-world applications. By enabling in situ digital interactions with non-instrumented analog objects, we can expand their utility (e.g., enabling a pot to double as a cooking timer), better synthesize their relevant information (e.g., comparing nutritional value), and overall enable richer interactivity and flexibility in everyday interactions. Here we present five example application scenarios from a broad application space we envision (see Figure 10) that highlight the value of XR-Objects.

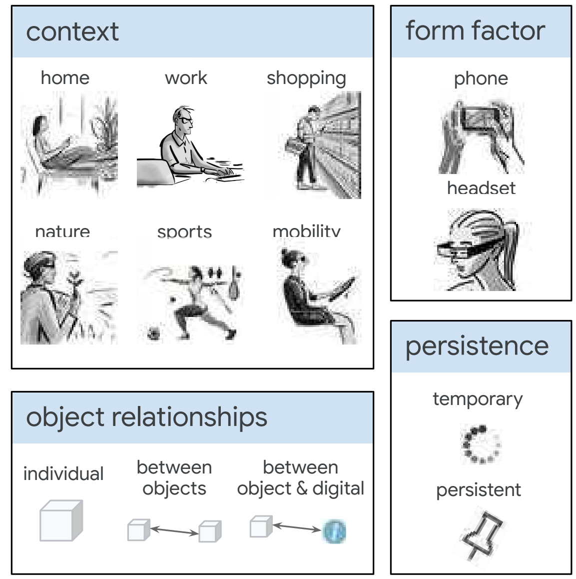

Figure 10: Envisioned application space for XR-Objects.

5.1 Discovery

XR-Objects enables users to discover new information about their surroundings by simply pointing their device at objects of interest. As shown in Figure 11, a user points their device at a vase containing different flowers and instantly receive information about each flower type, including its name, price, and care instructions. This on-demand, spatial discovery transforms everyday objects into avenues for understanding or appreciation of everyday objects that normally go unnoticed.

Figure 11: Example of discovery through MLLM interactions on top of XR-Objects.

5.2 Productivity

For productivity, XR-Objects could enhance physical documents with digital capabilities such as information retrieval and content anchoring. In Figure 12, a user reading a text book asks how it can be used to solve a particular type of equation, and anchors the response to the textbook for future reference. With added capabilities such as real-time optical character recognition (OCR) to digitize text, users could store and share digital copies of their physical documents for versioning and collaboration.

Figure 12: Interaction mock up of using XR-Objects for productivity on an HMD with direct touch interactions enabled.

5.3 Learning

XR-Objects offers immersive and interactive learning experiences by augmenting physical objects with contextual educational content. By pointing their device at an object, users can access relevant explanations, demonstrations, and quizzes that enhance their understanding of the subject matter.

As illustrated in Figure 13, XR-Objects can facilitate learning about healthy eating habits for children. A child can point their device at a fruit bowl and instantly see information about the different fruits, such as their names, nutritional values, and the vitamins they contain. This interactive learning experience can help children develop a deeper understanding of the importance of a balanced diet and the benefits of various fruits.

Figure 13: Example of anchoring contextual educational content over real-world objects using XR-Objects.

5.4 IoT Connectivity

XR-Objects's XR interface is complemented by a MLLM backend, unlike traditional IoT control interfaces that often limit interactions with devices to discrete apps. Through its object tracking, XR-Objects enables users to interact with their IoT devices in a spatial context, allowing for real-time visual feedback and control within their immediate environment.

For example, in Figure 14, the user controls their smart speaker using XR-Objects. They adjust the volume of the speaker using the touch UI or language commands utilizing the MLLM backend. Such connectivity scenarios expand beyond speakers and could also be enabled on thermostats, smart lights, and other edge IoT devices.

Figure 14: XR-Objects detects devices and provides custom UI elements for IoT control.

5.5 Cooking

Unlike traditional cooking aids, which rely on static recipes or digital screens detached from the cooking environment, XR-Objects integrates digital intelligence directly into the kitchen, making the cooking process informative and engaging (Figure 1). In our augmented cooking app, as the user places ingredients on the kitchen counter, our system recognizes each item and projects relevant information, such as freshness indicators, nutritional facts, and potential allergens, directly on the ingredients. Users can interact through voice commands or touch elements to ask about potential recipes or to compare ingredients. Using stopwatch or countdown timers, the system embeds the guidance into the cooking space itself. The user can share the final product with their contacts.

6 DISCUSSION

Our daily environment is ubiquitously augmented with various forms of annotations, from product packaging and price tags to traffic signs and personalized post-it reminders. This ubiquitous augmentation, a rudimentary form of augmented reality, is facilitated by the advent of language, writing, and scalable printing technologies. It serves as a means to asynchronously convey contextual information, albeit in a static and limited manner, requiring manual interpretation and application by the user. Although machine-readable markers like barcodes, QR codes, and NFC tags have simplified certain interactions, they fall short in offering dynamic, object-relevant actionable insights.

Augmenting Objects with Intelligence. Advancements in computer vision and LLMs now enable devices to not only recognize generic object categories but also distinguish between individual instances based on their spatial context. This unlocks the potential for personalized, instance-specific interactions to transform objects into intelligent entities with their own "memories" of past interactions. Accordingly, AOI has the potential to transition from the concept of smart tools to a reality where intelligence is an inherent characteristic of every object.

Context-Aware Interactions. Looking ahead, XR-Objects could evolve to become even more attuned to the user's context (e.g. current location and activity, object relationships, persistence of object and data, see Figure 10) to further customize actions and information based on historical interactions with objects in specific settings. For instance, viewing a food item's packaging in a store might trigger suggestions for understanding its allergens and nutritional content, while the same item at home could offer cooking instructions and allow the user to directly set a digital cooking timer to the appropriate duration. On certain smartphones, for instance, *Siri Suggestions*[10] already offer context-aware recommendations. The functionality of this sort of context-awareness could be extended using AR interfaces.

Leveraging Emerging Artificial General Intelligence (AGI). The integration of emerging AGIs [46], exemplified by models like *Gemini* or *GPT-4*, opens up new opportunities for autonomous problem-solving within XR environments. AGI's potential to dynamically generate user interfaces and visualizations in response to user queries could transform the way we interact with our physical world. For instance, the user could ask the system to visualize the nutritional values of a product in a pie chart, i.e., by generating the code to create the user interface and graphs on-the-fly without being pre-programmed to create the chart. Going a step further, one may imagine AGI-driven systems that not only respond to user prompts but also proactively offer assistance surfaced through a new action in the context menu, such as assembling a set of Lego blocks into a desired structure through real-time, augmented instructions.

Linking Realities. As we adopt these new innovations in AI and XR, we are likely to see significant changes in how we interact with the physical objects around us. We envision a future where physical items no longer need conventional labels or tags, relying instead on AOI for context and interaction. The merging of digital and physical worlds might bring about new connections, like *direct links between digital files and physical items.* This could start a new phase where the digital and physical worlds blend together smoothly, without clear boundaries.

7 CONCLUSION

In this paper, we introduced Augmented Object Intelligence (AOI), a novel paradigm that seamlessly integrates digital capabilities into physical objects through the use of XR-Objects. Our prototype system demonstrates the potential of AOI to transform how users interact with their surroundings by leveraging advancements in computer vision, spatial understanding, and Multimodal -LLM. The results of our user study show that XR-Objects significantly outperforms traditional multimodal AI interfaces, with participants completing tasks an average of 24% shorter time and reporting higher levels of satisfaction, ease of use, and perceived responsiveness. By enabling familiar, interactions with everyday objects through anchored AR content and natural language processing, XR-Objects paves the way for a future where the boundaries between the physical and digital worlds become increasingly blurred. As we continue to expand the capabilities of XR-Objects, we envision a wide range of applications spanning domains such as cooking, productivity, and IoT connectivity, ultimately leading to a more engaging, efficient, and immersive way of interacting with our surroundings.

[10] Siri Suggestions on iPhone: https://support.apple.com/guide/iphone/about-siri-suggestions-iph6f94af287/ios

REFERENCES

[1] Karan Ahuja, Sujeath Pareddy, 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 (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 189–196.

[2] David R Airth. 1993. Navigation in pop-up menus. In *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems (CHI '93)*. Association for Computing Machinery, New York, NY, USA, 115–116. event-place: Amsterdam, The Netherlands.

[3] Jonas Auda, Uwe Gruenefeld, Sarah Faltaous, Sven Mayer, and Stefan Schneegass. 2023. A scoping survey on cross-reality systems. *Comput. Surveys* 56, 4 (2023), 1–38.

[4] Nikola Banovic, Frank Chun Yat Li, David Dearman, Koji Yatani, and Khai N Truong. 2011. Design of unimanual multi-finger pie menu interaction. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces (ITS '11)*. Association for Computing Machinery, New York, NY, USA, 120–129. event-place: Kobe, Japan.

[5] Alan F Blackwell. 2006. The reification of metaphor as a design tool. *ACM Trans. Comput.-Hum. Interact.* 13, 4 (Dec. 2006), 490–530. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[6] Elise Bonnail, Wen-Jie Tseng, Mark Mcgill, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2023. Memory Manipulations in Extended Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg,Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 875, 20 pages.

[7] Arthur Caetano and Misha Sra. 2022. ARfy: A Pipeline for Adapting 3D Scenes to Augmented Reality. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–3.

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. arXiv:2209.06794 [cs.CV]

[9] Zhutian Chen, 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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[10] 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 (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 282–297. event-place: Virtual Event, USA.

[11] Yi Fei Cheng, Christoph Gebhardt, and Christian Holz. 2023. InteractionAdapt: Interaction-driven Workspace Adaptation for Situated Virtual Reality Environments. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA,USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 109, 14 pages.

[12] Subramanian Chidambaram, Hank Huang, Fengming He, Xun Qian, Ana M Villanueva, Thomas S Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2021. ProcessAR: An augmented reality-based tool to create in-situ procedural 2D/3D AR Instructions. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 234–249.

[13] Subramanian Chidambaram, Sai Swarup Reddy, Matthew Rumple, Ananya Ipsita, Ana Villanueva, Thomas Redick, Wolfgang Stuerzlinger, and Karthik Ramani. 2022. EditAR: A Digital Twin Authoring Environment for Creation of AR/VR and Video Instructions from a Single Demonstration. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Singapore, Singapore, 326–335.

[14] Neil Chulpongsatorn, Mille Skovhus Lunding, Nishan Soni, and Ryo Suzuki. 2023. Augmented Math: Authoring AR-Based Explorable Explanations by Augmenting Static Math Textbooks. (July 2023). _eprint: 2307.16112.

[15] Neil Chulpongsatorn, Wesley Willett, and Ryo Suzuki. 2023. HoloTouch: Interacting with Mixed Reality Visualizations Through Smartphone Proxies. (March 2023). _eprint: 2303.08916.

[16] Davison. 2003. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE, 1403–1410.

[17] Mustafa Doga Dogan, Raul Garcia-Martin, Patrick William Haertel, Jamison John O'Keefe, Ahmad Taka, Akarsh Aurora, Raul Sanchez-Reillo, and Stefanie Mueller. 2023. BrightMarker: 3D Printed Fluorescent Markers for Object Tracking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[18] Mustafa Doga Dogan, Alexa F. Siu, Jennifer Healey, Curtis Wigington, Chang Xiao, and Tong Sun. 2023. StandARone: Infrared-Watermarked Documents as Portable Containers of AR Interaction and Personalization. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7.

[19] Mustafa Doga Dogan, Ahmad Taka, Michael Lu, Yunyi Zhu, Akshat Kumar, Aakar Gupta, and Stefanie Mueller. 2022. InfraredTags: Embedding Invisible AR Markers and Barcodes Using Low-Cost, Infrared-Based 3D Printing and Imaging Tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3491102.3501951 Issue: Article 269 event-place: New Orleans, LA, USA.

[20] 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 *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–4.

[21] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. 2020. DepthLab: Real-time 3D interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 829–843.

[22] Andreas Rene Fender and Christian Holz. 2022. Causality-preserving Asynchronous Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA,USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 634, 15 pages.

[23] Redacted for anonymity during review process. 2024. Programmable Reality. *In Press* (2024).

[24] Eric J. Gonzalez, Khushman Patel, Karan Ahuja, and Mar Gonzalez-Franco. 2024. XDTK: A Cross-Device Toolkit for Input & Interaction in XR. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE.

[25] Mar Gonzalez-Franco and Andrea Colaco. 2024. Guidelines for Productivity in Virtual Reality. *ACM Interactions Magazine* (2024).

[26] Mar Gonzalez-Franco and Jaron Lanier. 2017. Model of illusions and virtual reality. *Frontiers in psychology* 8 (2017), 273943.

[27] Violet Yinuo Han, Hyunsung Cho, Kiyosu Maeda, Alexandra Ion, and David Lindlbauer. 2023. BlendMR: A Computational Method to Create Ambient Mixed Reality Interfaces. *Proceedings of the ACM on Human-Computer Interaction* 7, ISS (Nov. 2023), 436:217–436:241.

[28] Jeremy Hartmann, Christian Holz, Eyal Ofek, and Andrew D Wilson. 2019. Realitycheck: Blending virtual environments with situated physical reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[29] Fengming He, Xiyun Hu, Jingyu Shi, Xun Qian, Tianyi Wang, and Karthik Ramani. 2023. Ubi Edge: Authoring Edge-Based Opportunistic Tangible User Interfaces in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[30] Steven J. Henderson and Steven Feiner. 2008. Opportunistic controls: leveraging natural affordances as tangible user interfaces for augmented reality. In *Proceedings of the 2008 ACM symposium on Virtual reality software and technology (VRST '08)*. Association for Computing Machinery, New York, NY, USA, 211–218.

[31] 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 (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1957–1967.

[32] Valentin Heun, James Hobin, and Pattie Maes. 2013. Reality editor: programming smarter objects. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication (UbiComp '13 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 307–310.

[33] Rahul Jain, Jingyu Shi, Runlin Duan, Zhengzhe Zhu, Xun Qian, and Karthik Ramani. 2023. Ubi-TOUCH: Ubiquitous Tangible Object Utilization through Consistent Hand-object interaction in Augmented Reality. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–18.

[34] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.

[35] Bruce M King, Patrick J Rosopa, and Edward W Minium. 2018. *Statistical reasoning in the behavioral sciences*. John Wiley & Sons.

[36] Yoshiki Kudo, Anthony Tang, Kazuyuki Fujita, Isamu Endo, Kazuki Takashima, and Yoshifumi Kitamura. 2021. Towards balancing VR immersion and bystander awareness. *Proceedings of the ACM on Human-Computer Interaction* 5, ISS (2021), 1–22.

[37] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746* (2022).

[38] Julian Lepinski, Eric Akaoka, and Roel Vertegaal. 2009. Context menus for the real world: the stick-anywhere computer. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 3499–3500. event-place: Boston, MA, USA.

[39] Zhen Li, Michelle Annett, Ken Hinckley, Karan Singh, and Daniel Wigdor. 2019. HoloDoc: Enabling Mixed Reality Workspaces that Harness Physical and Digital Content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300917 Issue: Paper 687 event-place: Glasgow, Scotland Uk.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs].

[41] 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 (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 147–160.

[42] David Lindlbauer, Jens Emil Grønbæk, Morten Birk, Kim Halskov, Marc Alexa, and Jörg Müller. 2016. Combining Shape-Changing Interfaces and Spatial Augmented Reality Enables Extended Object Appearance. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 791–802.

[43] David Lindlbauer and Andy D Wilson. 2018. Remixed reality: Manipulating space and time in augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[45] Kyzyl Monteiro, Ritik Vatsal, Neil Chulpongsatorn, Aman Parnami, and Ryo Suzuki. 2023. Teachable Reality: Prototyping Tangible Augmented Reality with Everyday Objects by Leveraging Interactive Machine Teaching. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15.

[46] Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462 [cs.AI]

[47] Majid Pourmemar and Charalambos Poullis. 2019. Visualizing and Interacting with Hierarchical Menus in Immersive Augmented Reality. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9.

[48] Shwetha Rajaram and Michael Nebeling. 2022. Paper Trail: An Immersive Authoring System for Augmented Reality Instructional Experiences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16.

[49] Wendy A Rogers, Tracy L Mitzner, and Michael T Bixter. 2020. Understanding the potential of technology to support enhanced activities of daily living (EADLs). *Gerontechnology* 19, 2 (2020).

[50] Krystian Samp and Stefan Decker. 2010. Supporting menu design with radial layouts. In *Proceedings of the International Conference on Advanced Visual Interfaces*. ACM, Roma Italy, 155–162.

[51] Maximilian Speicher, Brian D. Hall, and Michael Nebeling. 2019. What is Mixed Reality?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–15.

[52] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. arXiv:2310.12953 [cs].

[53] Ryo Suzuki, Mar Gonzalez-Franco, Misha Sra, and David Lindlbauer. 2023. XR and AI: AI-Enabled Virtual, Augmented, and Mixed Reality. In *Adjunct Proceedings*

of the 36th Annual ACM Symposium on User Interface Software and Technology. ACM, San Francisco CA USA, 1–3.

[54] Ryo Suzuki, Rubaiat Habib Kazi, Li-yi Wei, Stephen DiVerdi, Wilmot 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*. ACM, Virtual Event USA, 166–181.

[55] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. 2023. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469* (2023).

[56] Jenifer Tidwell, Charles Brewer, and Aynne Valencia. 2020. *Designing Interfaces: Patterns for Effective Interaction Design* (3 ed.). O'Reilly Media. https://www.amazon.com/Designing-Interfaces-Patterns-Effective-Interaction/dp/1492051969

[57] Wen-Jie Tseng. 2023. Understanding Physical Breakdowns in Virtual Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg,Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 506, 5 pages.

[58] Wen-Jie Tseng, Elise Bonnail, Mark McGill, Mohamed Khamis, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2022. The Dark Side of Perceptual Manipulations in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA,USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 612, 15 pages.

[59] Julien Valentin, Adarsh Kowdle, Jonathan T Barron, Neal Wadhwa, Max Dzitsiuk, Michael Schoenberg, Vivek Verma, Ambrus Csaszar, Eric Turner, Ivan Dryanovski, et al. 2018. Depth from motion for smartphone AR. *ACM Transactions on Graphics (ToG)* 37, 6 (2018), 1–19.

[60] Tingting Wu, Alexander J Dufford, Laura J Egan, Melissa-Ann Mackie, Cong Chen, Changhe Yuan, Chao Chen, Xiaobo Li, Xun Liu, Patrick R Hof, and Jin Fan. 2017. Hick–Hyman Law is Mediated by the Cognitive Control Network in the Brain. *Cerebral Cortex* 28, 7 (05 2017), 2267–2282.

[61] Hanyu Xiang, Qin Zou, Muhammad Ali Nawaz, Xianfeng Huang, Fan Zhang, and Hongkai Yu. 2023. Deep learning for image inpainting: A survey. *Pattern Recognition* 134 (2023), 109046.

[62] Chang Xiao, Ryan Rossi, and Eunyee Koh. 2022. iMarker: Instant and True-to-scale AR with Invisible Markers. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–3.

[63] Jackie (Junrui) Yang and James A. Landay. 2019. InfoLED: Augmenting LED Indicator Lights for Device Positioning and Communication. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*. Association for Computing Machinery, New York, NY, USA, 175–187.

[64] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A Survey on Multimodal Large Language Models. arXiv:2306.13549 [cs.CV]

[65] Yuguang Zeng and Jingyuan Zhang. 2014. Multiple user context menus for large displays. In *Proceedings of the 2014 ACM Southeast Regional Conference (ACM SE '14)*. Association for Computing Machinery, New York, NY, USA, 1–4. Issue: Article 44 event-place: Kennesaw, Georgia.

[66] Zhengzhe Zhu, Ziyi Liu, Tianyi Wang, Youyou Zhang, Xun Qian, Pashin Farsak Raja, Ana Villanueva, and Karthik Ramani. 2022. MechARspace: An Authoring System Enabling Bidirectional Binding of Augmented Reality with Toys in Real-time. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Association for Computing Machinery, New York, NY, USA, 1–16.

A STUDY MATERIALS

A.1 HALIE Survey

Following the completion of all tasks in a given condition (XR-Objects , Chatbot), participants rated their agreement with the following statements on a 5-point *Likert* scale, and provided open ended answers to the questions denoted with (*O):

- H **(Helpfulness)** Independent of its fluency, the AI Tool was helpful for completing my task.
- *O1 **(Helpfulness)** What kinds of aspects did you find helpful or not helpful and why? (Give a concrete example if possible.)
- J **(Enjoyment)** It was enjoyable using the AI tool to accomplish the tasks.
- S **(Satisfaction)** Independent of its fluency, I am satisfied with *how* the AI tool provided its answers.
- R **(Responsiveness)** Independent of its fluency, I found the AI tool to be a responsive system.
- E **(Ease)** Overall, it was easy to interact with the AI Tool and accomplish the tasks.
- ER **(Ease)** Getting information about an object was easy using the AI tool.
- E2 **(Ease)** Comparing two objects was easy using the AI tool.
- EN **(Ease)** Comparing more than two objects was easy using the AI tool.
- *O2 **(Change)** Did you change how you chose to interact with the AI Tool over the course of the task? If so, how?
- *O3 **(Description)** What adjectives would you use to describe the AI Tool?
- *O4 **(Compare)** How did this interaction compare to your regular in-person shopping and at-home task experiences?

A.2 Form Factor Survey

For each question, participants selected either Chatbot or XR-Objects. They did this questionnaire twice, once we were asking participants to think XR-Objectswas going to run on a Phone, the second time thinking XR-Objectswould run on a headset form factor. This questionnaire included the adapted HALIE questions (H, J, S, R, ER, E2, EN) as well as an additional set of questions detailed below:

- EC **(Ease Communications)** Sending a message about one of the grocery items would be easier on headset/phone using:
- ET **(Ease Timer)** Setting a timer would be easier on headset/phone using:
- EN **(Ease Note)** Creating a note (e.g., reminder to buy more juice) would be easier on a headset/phone using:
- IS **(Improved Shopping)** Which AI Tool running on a headset/phone would represent a better change compared to your current experience with (in-person) shopping?
- P **(Preference)** Overall, which AI tool would you prefer on headset/phone?