PaperToPlace: Transforming Instruction Documents into Spatialized and Context-Aware Mixed Reality Experiences

Chen Chen*
University of California San Diego
La Jolla, CA, United States
chenchen@ucsd.edu

Cuong Nguyen
Adobe Research
San Francisco, CA, United States
cunguyen@adobe.com

Jane Hoffswell
Adobe Research
Seattle, WA, United States
jhoffs@adobe.com

Jennifer Healey
Adobe Research
San Jose, CA, United States
jehealey@adobe.com

Trung Bui
Adobe Research
San Jose, CA, United States
bui@adobe.com

Nadir Weibel
University of California San Diego
La Jolla, CA, United States
weibel@ucsd.edu
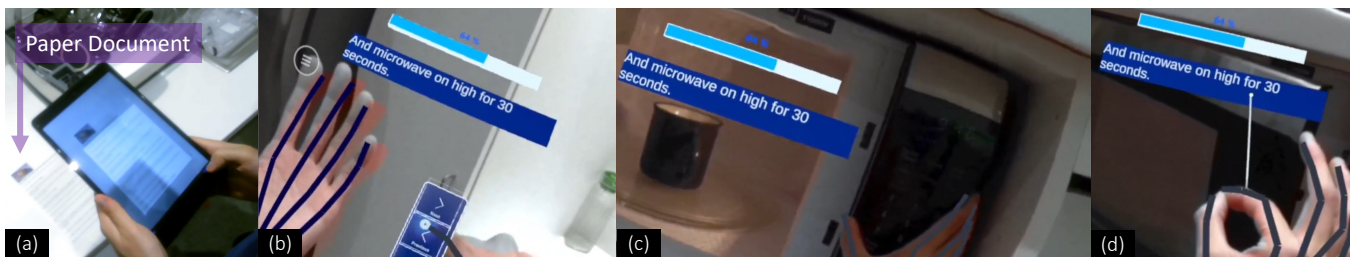
Figure 1: Overview of PaperToPlace: (a) The author creates an MR experience by taking a snapshot of a paper document, with an optional ML-supported pipeline for associating key objects with each instruction step; (b) The consumer can browse the spatialized instruction steps using a hand menu; (c) The step is placed at an optimal location to minimize context switching and prevent occlusion of important interaction areas (*e.g.,* not occluding the touchpad while setting the time on a microwave); (d) The consumer can "pinch-and-drag" the step to refine the system placement. Steps (b - d) show the first-person MR view.

ABSTRACT

While paper instructions are a mainstream medium for sharing knowledge, consuming such instructions and translating them into activities can be inefficient due to the lack of connectivity with the physical environment. We propose PaperToPlace, a novel workflow comprising an *authoring pipeline*, which allows the authors to rapidly transform and spatialize existing paper instructions into an MR experience, and a *consumption pipeline*, which computationally places each instruction step at an optimal location that is easy to read and does not occlude key interaction areas. Our evaluation of the authoring pipeline with 12 participants demonstrates the usability of our workflow and the effectiveness of using a machine learning based approach to help extract the spatial locations associated with each step. A second within-subjects study with another 12 participants demonstrates the merits of our consumption pipeline to reduce context-switching effort by delivering individual segmented instruction steps and offering hands-free affordances.

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**.

KEYWORDS

Mixed Reality (MR), Instruction Document, Tutorial and Help Systems, Context–Aware MR

1 INTRODUCTION

Paper-based instructions are common for knowledge sharing. Such instructions are often related to tasks that require users to interact with multiple objects spatially distributed in an environment. For example, when following a recipe, a user may need to interact with multiple kitchen appliances like the cooktop, fridge, and microwave. When following a safety manual, a compliance manager may need to interact with various machines on the factory floor.

However, performing a task while consuming instructions can be tedious as the text is typically *disassociated* from the user's physical environment. Thus, a user has to balance reading the instructions, figuring out what they mean in the environment, and performing the task, which can be cognitively demanding [37]. For example,

when frying a piece of steak while following a cookbook, one needs to frequently switch between the cookbook and the pan to check the searing technique, temperature, *etc.* This switch can be costly if the user places the cookbook somewhere peripheral so that it does not obstruct the task area. The user might end up spending more time trying to navigate the text and environment, than performing the task. This problem is made worse if the user forgets some important information like temperature or duration, and has to repeatedly come back to the instruction to double check.

Consumer Augmented Reality (AR) and Mixed Reality (MR)[1] offer a unique opportunity to address this document-activity disassociation by overlaying digital elements on top of the environment. These approaches are becoming more accessible, and studies have demonstrated their potentials for training workers to conduct tasks that are spatial in nature [44]. While prior works investigated the affordances of virtual guidance for conducting spatial tasks [48], and how to integrate such guidance in MR [33], they have not explored how document contents and their associated consumption experience could be constructed in MR. To that end, Microsoft Dynamic 365 Guides [12] is an industry solution to help enterprise users manually create instructions and anchor them in an MR experience.

Across these MR instruction systems, the placement of the virtual instructions is often decided by the authors, and cannot be dynamically adapted to real-world contexts. This assumption could lead to undesirable experiences in both the consumption phase and the authoring phase. For the consumption phase, a static MR instruction could be mistakenly placed at a location too far from the user's task, at a distance that is difficult to read, or at a position that occludes key objects that the user is interacting with. For authoring, the author has to spend time associating and placing an instruction with its corresponding physical object. This process is time consuming and has to be repeated for every new set of instructions, even though the physical layout of objects might not change much over time. While some prior works, *e.g.,* FLARE [41], showed the usefulness of creating a persistent AR layout, real-world instructional activities are frequently changing (*e.g.,* users might move from one place to another depending on the procedural step in the instructions), causing such a static layout to be infeasible.

We propose PaperToPlace, a novel end-to-end workflow that transforms paper instructions into a context-aware instructional MR experience by *segmenting* monolithic documents; *associating* instruction steps with real-world anchoring objects; and optimally *placing* the virtual instruction steps so that they are easy to read and revisit while completing the tasks. To realize this goal, PaperToPlace consists of an *authoring* and a *consumption* pipeline. With the authoring pipeline, the author can simply take a snapshot of the paper document by leveraging a mobile camera (Fig. 1a). Our system then segments the text in the document into individual instruction steps. The author can manually edit these segments, and *associate* each step with the spatial location where the relevant activities will occur. To help with this association task, we designed a machine learning (ML) approach, where a fine-tuned language model was used to suggest the relevant spatial locations to the author. Our

consumption pipeline is designed to render these instruction steps and place them in the associated spatial locations. To ensure users could easily consume the spatialized instruction, the placement of each step in MR is optimized using a probabilistic optimization approach based on pre-created environmental models and the tracked gaze and hands. Fig. 1c shows an example of a spatialized instruction step, where the step "*And microwave on high for 30 seconds*" is tagged with "*microwave*" and is optimally rendered in front of the user while not occluding their view as they set the heating time.

We prototyped PaperToPlace on Meta Quest Pro [11], and conducted two within-subjects studies to evaluate the authoring pipeline with 12 participants, and the consumption pipeline with another 12 participants. We demonstrated the usability of our authoring pipeline, and the effectiveness of using an ML-based approach to help the authors extract the spatial location associated with each step. We then illustrated the effectiveness of the consumption pipeline for reducing context-switching effort, delivering the segmented instruction steps, and offering hands-free affordances.

With the assumption that the spatial profiles (*i.e.,* the environmental geometry and associated semantic labels, see Sec. 4.1) are available, we contribute the design and evaluations of:

- **An authoring pipeline** that allows users to transform paper instructions into a spatialized MR experience;
- **A consumption pipeline** that can computationally place the virtual instruction steps in the optimal place without either occluding the user's view or leading to large degrees of context switching.

2 RELATED WORK

This paper is motivated by prior work on incorporating instruction experiences into MR and designing context-aware MR experiences.

2.1 Integrating Instruction Experiences into MR

MR has been widely used for augmenting document consumption experiences [29, 30, 59]. Augmenting instructional documents, however, is still challenging due to the need to connect and integrate with real-world scenes and activities [42, 64]. Many prior works have explored the use of MR to augment a procedural instruction experience — an important asynchronous collaboration task. For example, ProcessAR [33] proposed *in-situ* procedural AR instructions that could be rapidly created by experts, and used to teach novices through spatial and temporal demonstrations (*e.g.,* information about how to move a tool in the temporal domain and orient it in the spatial domain). However, the placement of textual instructions was not explored. CAPturAR [68] introduced a MR tool that helps users rapidly author context-aware applications, by referring to recorded activities. Commercial tools such as Microsoft Dynamic 365 Guides [12] enables experts to author a MR instruction experience by enacting the guidance, placing the instruction in the designated space, and recording the tool operations.

Although these works explored the design of MR-based instruction experiences, existing paper instructions are usually left behind, resulting in unnecessary time and effort to redesign a usable instruction workflow. Additionally, existing MR instruction experiences are often not able to dynamically adapt to the changing environmental context of real-world activities (*e.g.,* [12, 33]), causing user

---

[1] While AR and MR are frequently interchangeable in literature, our contribution lies under the MR paradigm that focus on interactivity and context awareness [66].

frustration when virtual graphics occlude interaction tasks. Paper-ToPlace is novel in that it supports reusing existing paper documents that are designed by professional writers in a reader-centered way [69], and can transform such documents into a spatialized and context-aware MR experience that is adapted to both the user's needs and the environmental characteristics.

2.2 Designing Context-Aware MR Experiences

Context-aware MR systems aim to show *"the 'right' information, at the 'right' time, in the 'right' place, in the 'right' way"* [40], which requires understanding both human and environmental contexts.

Prior research explored various computational approaches to realize this goal. For example, Lindlbauer *et al.* [53] used the real-time cognitive load, estimated by pupil dilation, to decide when and where the application should be shown, as well as how much information should be delivered (*i.e.,* level of detail) in an MR system. Lang *et al.* [51] used simulated annealing to place virtual agents by considering key anchoring surfaces in the environment identified by pre-trained mask R-CNN. Liang *et al.* [52] used a similar approach to build a scene-aware virtual pet which could behave naturally in the real-world (*e.g.,* respond to a food bowl). Yu *et al.* [71] proposed an interactive and context-aware furniture recommendation MR system by considering the real-world scene (*i.e.,* spatial context) and the learnt furniture compatibility in a latent space (*i.e.,* category context). ScalAR enabled designers to author a semantically adaptive AR experience [60]. Liu *et al.* [54] attempt to generate suggestions for the arrangement of work surfaces in HoloLens, by capturing users' habitual behaviors of interacting with objects on the work surface. Similar to our work, SemanticAdapt [32] used an optimization approach to automatically adapt MR layouts between different environments by considering the virtual-physical semantic connections. However, the target applications were only related to information consumption (*e.g.,* consuming news feeds) and did not consider those related to real-world activities.

Inspired by this prior work, PaperToPlace demonstrates a novel consumption pipeline that leverages a similar computational approach to analyze the tracked gaze and hand position, as well as the anchoring surfaces of the key objects in a target environment.

3 PRELIMINARY NEEDS-FINDING STUDY

To understand the pain-points for consuming existing paper instructions (*i.e., monolithic, non*-segmented, *non*-spatialized, and *non*-context-aware), we conducted a needs-finding study, rooted in participant observations [23, 67] and semi-structured interviews.

3.1 Participants, Tasks, and Procedure

We recruited four participants (age, $M = 24.75$, $SD = 2.87$, two females, two males). Participants were required to complete the designated task using paper instructions. Specifically, PP1 and PP4 were required to make coffee with a coffee machine by following the user manual [7] (Fig. 2a). PP2 and PP3 were required to make a chocolate microwave cup cake from an existing recipe [14] (Fig. 2b). These tasks were chosen since they are common activities in an office kitchen; require participants to read the textual instructions; and could be completed in a reasonable time for an unpaid study.
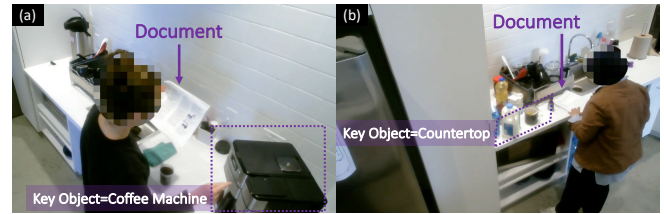
Figure 2: Preliminary needs-finding tasks. (a) Making a cup of coffee using a coffee machine (PP1). (b) Making a chocolate cake in a mug with a microwave (PP3).

We also used the existing instructions [7, 14] created by professional writers to minimize the impact of non-professional writing styles. All participants reported little (PP3, PP4) to no (PP1, PP2) prior knowledge for the designated task. Next, we conducted a semi-structured interview, focusing on *"what are the pain-points while performing the designated tasks using the given paper instruction, and why?"* Finally, we brainstormed potential designs of a MR experience for consuming instruction documents. We explained the concept of MR for PP3 and PP4 who were not familiar with it. Participants were encouraged to sketch their imagined design on an iPad Canvas. The study took on average 30 min ($SD = 3.25$ min).

3.2 Findings

Overall, we identified three pain-points through the study.

The overwhelming amount of information or lack of necessary details in the instructions can impact the usability. During the semi-structured interviews, two participants (PP1, PP3) pointed out that the sometimes overwhelming amount of information and irrelevant content could be distracting. For example: *"there was a lot of information on the document. And it wasn't easy for me to know where and what information I should be looking at"* (PP1) and *"the first setback was too much information"* (PP3). To handle the potentially overwhelming amount of information, PP2 first skimmed the document in search of relevant content: *"I am first attracted to see where the bullet points are. [...] And then if I just skim through the first two or three points, I understand that this is not relevant. So I'm just skipping those sections completely."* While designing the MR experience, PP1 incorporated such insights into her design (Fig. 3c) and commented: *"I would rather the MR just gives me one small step every time. For example, I am making coffee and reaching the step two, and in the virtual instruction, it will say like coffee making step two, and then here will be just a instruction with just a few sentences."*

On the other hand, PP2 and PP4 believed the lack of details for certain steps could impact their ability to perform the sub-task. For example: *"I've never cracked an egg so I don't know how to do it. [...] If it is some things that I've never actually done [and the instruction document does not tell me how], I might actually be confused"* (PP2). In essence, different users likely require different levels of information based on their prior experience with the intended task.

There are missing connections between the instruction step and real-world activities. All participants identified a need for establishing spatial connections between instructions and real-world objects or activities. For example, during the design phase of the study, PP2 emphasized that *"having [a virtual] arrow [in MR] that can help me and connect me to the object is helpful."*
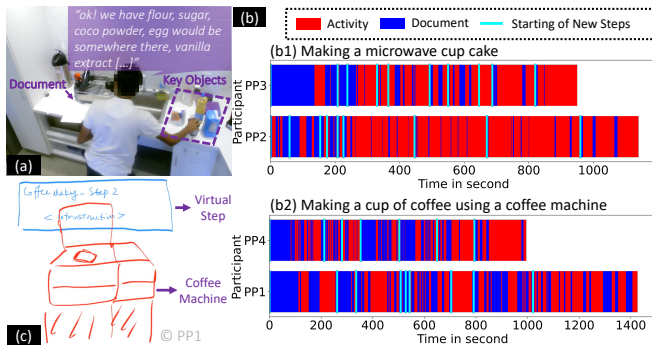
Figure 3: Preliminary needs-finding study results. (a) Example context switching while PP2 was attempting to map instructions with real-world objects. (b) The annotated timestamps showing participants' current focus as either the document or real-world activities. (c) PP1's design for an instructional MR experience while using a coffee machine.

On the other hand, most participants suggested that overwhelming spatial guidance might be unnecessary and could lead to visual disturbance, similar to their concerns about the potentially overwhelming amount of information already in some instructional documents. Reflecting further on PP2's suggestion to include virtual arrows, he highlighted some potential downsides of this approach, noting that *"[having useless spatial indicators] is going to be overloaded. I know these basic things, and I don't need pointers to see like a spoon or a mug."* PP1's design sketches implied an alternative approach to establish connections between the instruction step and real-world objects by placing the virtual step close to the coffee machine, without occluding the user's view (Fig. 3c).

Frequent context-switching between the instruction document and real-world activities should be minimized. By analyzing the video recording, we found that users perform frequent context-switching between documents and real-world activities. For example, while PP2 was conducting the task, he naturally commented: *"ok! we have flour, sugar, coco powder, egg would be somewhere there, vanilla extract [...]"*, thereby demonstrating frequent context-switching between instructions and real-world objects as he checked off items from the recipe list (Fig. 3a). Participants generally used two types of strategies to minimize context-switching: *holding* and *switching to* the document while performing the task. Specifically, we found that PP1 tended to hold the documents while performing tasks, although occasionally this approached was inconvenient for steps requiring two hands (Fig. 2a). In contrast, other participants tended to place the document on the countertop while reading the content, and move the document to a new place when switching to other steps (Fig. 2b). To understand how participants used the documents to perform individual steps, we manually labelled the timestamp of the recorded video while participants were switching between documents and real-world activities, or otherwise. Through this process, we observed that the participants highly relied on the documents to perform tasks. Specifically, while participants were attempting to complete a particular step, the document was still frequently referenced even though the participants had already read it at the beginning of each step (Fig. 3b).

3.3 Design Considerations

We identified three fundamental <u>D</u>esign <u>C</u>onsiderations (DCs) by analyzing the data from our preliminary needs-finding process.

[DC1] Only delivering the segmented instruction could enhancing information consumption experience. We show that the participants expect to consume relevant information corresponding to their current activities, yet existing paper documents usually deliver all information to users at the same time. An improved MR instruction consumption experience could create novel and flexible ways to segment the document, such that only relevant information is delivered to users for each associated step.

[DC2] Optimally placing instruction texts next to the areas of interactive activities might be helpful for the intermittent and repetitive information consumption experience. While participants generally read the entirety of an instruction step before performing the corresponding actions, the instruction step may need to be consumed *repeatably* (Fig. 3b). Therefore, the placement of the virtual instruction step in MR should consider the spatial location where the relevant task would occur.

[DC3] The right level of spatial guidance could help users associate instructions with spatialized key objects. While few existing works (*e.g.,* [48, 49]) suggested the usefulness of spatial guidance for MR-based instructional experiences, our participants emphasized the importance of *moderate* spatial guidance, with limited visual disruptions. Thus, usable spatial guidance, without causing overwhelming visual disturbance, should be provided at the beginning of each instruction step.
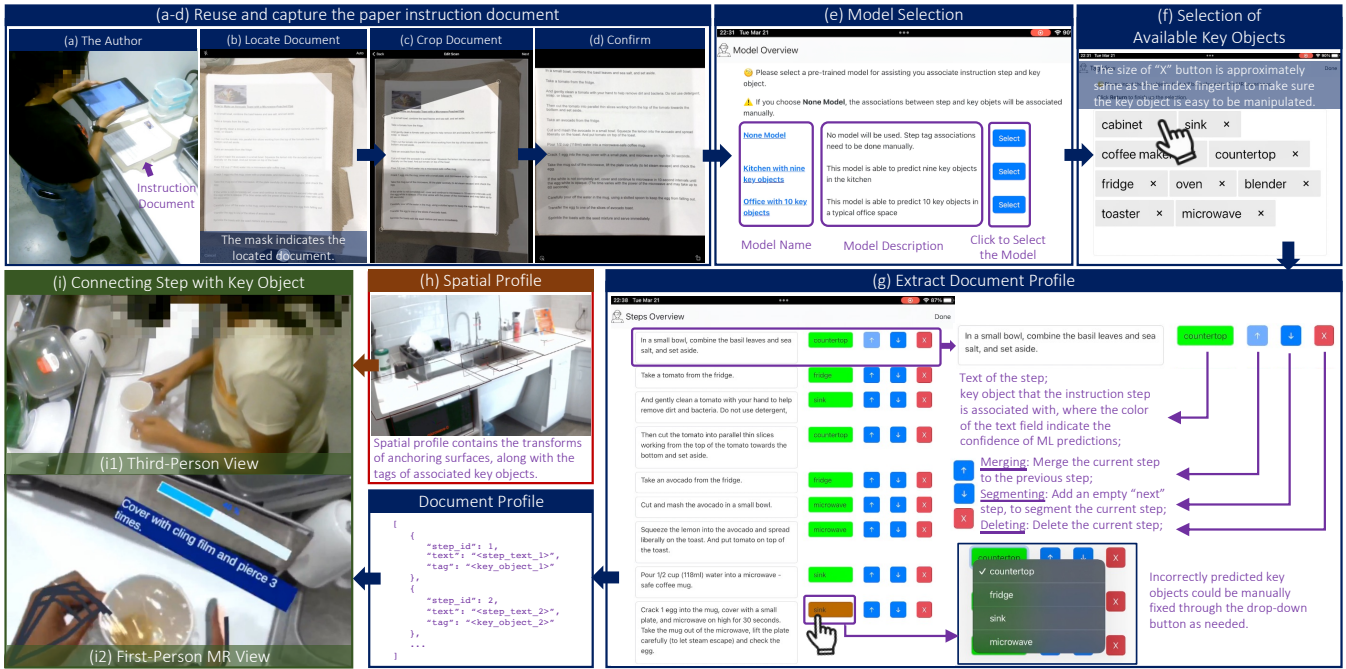
4 PAPERTOPLACE SYSTEM OVERVIEW

Based on Sec. 3.2, we designed PaperToPlace (Fig. 4), comprising of two pipelines: **(1) an authoring pipeline** for an **author** to rapidly and easily create a spatialized MR instruction experience from existing paper-based instructions (Fig. 4a - g, Sec. 5), and **(2) a consumption pipeline** for enabling a **consumer** to explore context-aware, spatialized instruction steps in MR (Fig. 4i, Sec. 6). While we use cooking tasks in an office kitchen as a running example for the design and evaluations, our approach could be transferred to other types of instruction documents.

4.1 Assumptions and System Walkthrough

We consider an **environment** to be a typical workspace (*e.g.,* the kitchen) for supporting a procedural **task** (*e.g.,* baking a cake). Each environment contains multiple physical **key objects**, which are defined as the important, stationary objects that are usually attached to the environment permanently (*e.g.,* the fridge and microwave). We did not consider non-stationary objects (*e.g.,* a mug) due to the lack of support for real-time arbitrary object tracking with Quest Pro [11]. Each key object contains one or multiple **anchoring surfaces**, which are virtual surface(s) that describe the approximated geometry of the objects. We use these surfaces to determine the placement of instruction in MR.

We also define a **spatial profile** as a collection of the labels of these key objects and their anchoring surface(s). We assume that a spatial profile could be created offline, either through automatic geometry processing (*e.g.,* [6]) or manual means (Fig. 5). Example

Figure 4: PaperToPlace system overview. We assume a spatial profile (h, **red** block) was pre-created. The author uses the authoring pipeline (a - g, **blue** blocks) to extract the document profile for the MR experience. With consuming pipeline (i, **green** block), the instruction steps are displayed based on the environmental (loaded via the spatial profile) and user's contexts.

environment where spatial profiles could be created in advance include office kitchen, a rental house, and a factory. To realize this assumption, we implemented a MR interface in the Quest Pro [11] that allow each anchoring surface (represented by a 2D plane) and the associated key object could be easily declared (Fig. 5a). The spatial anchor APIs [16] were used to ensure the declared anchoring surfaces are persistent in the environment (Fig. 5b).

With these assumptions, the goals of the two pipelines of Paper-ToPlace are described as below:

• **Authoring Pipeline.** Given an existing paper-based instruction document, the author would first segment the document content into smaller steps, each of which will only be associated with one key object. For each step, the author needs to identify the **metadata**, including: **(i)** the text of each instruction step, and **(ii)** the key object that the step should be anchored on. Together, such metadata makes up our **document profile** — the essential elements to recreate a spatialized MR experience from existing paper documents.
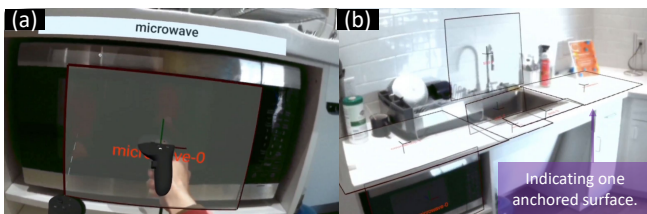
Figure 5: (a) Creation of an anchoring surface, visualized as a semi-transparent mask, using touch controllers. (b) Examples of anchoring surfaces in our experimental kitchen. Both scenes were captured as first-person MR views.

• **Consumption Pipeline.** While consuming the document in MR, the instruction steps float in mid-air and will be optimally attached to one of the anchoring surfaces of the associated key object, based on **user context**—real-time user interactions data in MR such as the tracked eye gaze and hand joints. For example, the step *"boil a cup of water in the microwave for 5 min"* should be attached to one of the anchoring surfaces of the *"microwave"* key object and not impact the user's interactions.

4.2 Application Scenarios

We consider two user roles, where the **author** use the authoring pipeline to rapidly create an instruction MR experience and the **consumer** use the consumption pipeline to consume the authored MR experience while completing tasks. We target on colocated and asynchronous collaborations [47], where the instructions are authored and consumed in same environment. Specific application scenarios are described below.

Author and Consumer are Different Users. For a specific procedural task, PaperToPlace could be used to facilitate asynchronous collaborations between experts and novices. For example, company administrators could use PaperToPlace for training new employees to use the provided facilities (*e.g.,* coffee machines and fridges in the shared office kitchen). Chidambaram *et al.* [33] also demonstrated the usefulness of using a similar instruction MR experience to teach novices assembly mechanics.

Author and Consumer are the Same User. While we differentiate two user roles, it is possible that the author and consumer are the *same* user. As the cognitive processes for consuming instructions usually occur in working memory, which is constrained by both

time and processing capacity, it is often necessary for one to revisit the procedures repeatably for the *same task* [21, 42]. For example, while cooking the same meal, it is common for the user to refer back to the cookbook each time upon starting a new step. Fig. 3b confirmed such patterns, where participants repeatedly refer to the instructions while completing a specific step. In this scenario, the user could author a personalized and spatialized MR experience, which s/he could use repeatably when preparing the same meal in the future; for example, a user could customize a recipe to take into account his/her preferences for spice level.

5 AUTHORING PIPELINE

The authoring pipeline extracts the document profile from an existing paper document to create an MR experience *rapidly* and *easily*.

5.1 Document Capture and Parsing

One question for document reuse is *how to enable users to rapidly capture and extract the document profile from an existing instruction document*? Inspired by mobile applications that allow users to capture and analyze scanned documents (*e.g.,* Adobe Scan [3] and Tab [17]), we similarly enable authors to simply take a snapshot of the instruction document to generate a document profile (Fig. 4a - d).

The author can then adjust the scanned region to crop out unnecessary components (*e.g.,* titles, *etc.*) as needed (Fig. 4c). Paper-ToPlace then leverages OCR services by Google Vision API [13] to parse the scanned image into machine readable text, due to its ability to extract paragraph structure in the parsed text using full text annotations [9]. By default, we segment each paragraph as one step in the instructions. However, the author can re-segment the steps and fix errors in a dedicated mobile interface (Fig. 4g).

5.2 Selecting the Model and Key Objects

To extract the document profile, we designed a manual and ML-assisted approach to help authors rapidly and easily associate key objects with each step. Our ML-assisted approach leverages a pre-trained language model for a specific environment to predict the key object that is associated with each step. After transforming the existing paper document into machine readable text, the authors need to select the model for the target environment (Fig. 4e). For environments without a pre-trained model, the manual approach enables authors to manually extract the metadata of each step.

The author then selects the key objects that exist in the target workspace (*i.e.,* the set of *available* key objects) (Fig. 4f). First, this step ensures the key objects contained in the extracted document profile aligns with the spatial profile of the intended environment. For example, a cooking instruction step such as *"boil a cup of water"* could be executed either in a typical household kitchen on a *cooktop*, or in an office kitchen that only provides a *microwave*. Second, setting the available key objects also provides prior knowledge to help increase the accuracy while predicting key object associations. For example, if an office kitchen only has a *microwave*, the key object for *"boil a cup of water"* should not be predicted as *oven*, even though *oven* is a possible label for the pre-trained model (Sec. 5.3).

(a)   (b)
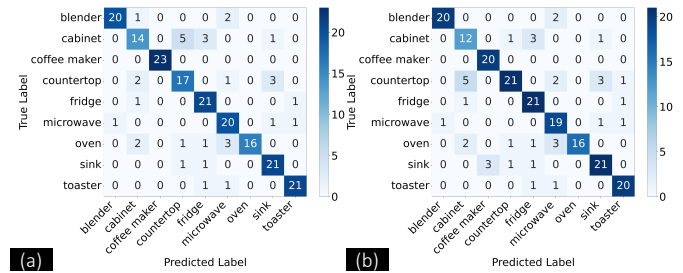
Figure 6: Confusion matrices of the fine-tuned BERT model, with the ground truth generated by rule-based method (a) and manually labeling (b).

5.3 Creating Document Profile

Creating a document profile requires two types of metadata:

(1) The text of each procedural step. While by default we consider and segment each sentence as one instruction step (Sec. 5.1), the author could overwrite the system segmented results by *segmenting*, *merging* and *deleting* specific step(s) (Fig. 4h). When a specific step is modified or a new merged step is generated, the associated key objects will be re-predicted (if the ML supported mode is used). Although some instruction steps might be associated with multiple or no key objects, such flexibility allows the author to split or merge the target step(s). In response to **[DC1]**, additional flexibility is also provided for the author to modify the generated text of each step to ensure that the right information with right level of details could be delivered to the consumers.

(2) The key object that the step is associated with. The authors can use either a manual or ML-assisted approach to determine the key object associated with each step. Fig. 4g shows a dedicated interface with segmented instructions, where the authors can use the drop downs to select the associated key objects, with the color scale indicating the confidence of our ML predictions (if applicable).

While manually assigning each step to a key object could work robustly, predicting key objects using a ML-assisted approach that requires a pre-trained model is challenging, due to the needs for a dataset and ground truth labels. Creating such a dataset that could be generalized to *all* procedural instructions is not realistic, and labeling each step with a ground truth key object is also difficult and time consuming. Instead of preparing such dataset for *all* instruction documents, we chose to focus on domain-specific dataset that is publicly available. Alternatively, the dataset could be created via vendors or crowdsourcing.

We describe our methods for generating such a pre-trained model below. Although our running example is based on cooking instructions, the overall approach could be transferred to other type of instructions documents, provided the unlabeled dataset is available.

Dataset: We used RecipeNLG [22] for training purposes, which contains more than 2.2M cooking recipes where each recipe includes multiple ordered instruction steps. The process of aggregating all steps from all recipes yielded an unlabelled dataset with 19.5M steps, with an average of 11.54 (SD = 7.13) words per step.

Rule-Based Labelling of Training Instruction Steps: Instead of manually labelling each step, we used a rule-based approach to label steps that contain the *exact* words of the predefined key

objects. For example, we label the step *"boiling a cup of water in the microwave for 5 min"* as *"microwave"*, yet the step *"boiling a cup of water for 5 min"* will not be labeled, and thus will not be included as part of our training dataset. We iteratively selected nine key objects that exist in a typical kitchen: *blender, cabinet, coffee maker, countertop, fridge, microwave, oven, sink, toaster*. Such selections also ensures a reasonable amount of instruction steps for subsequent model fine-tuning purposes. We generated a dataset where each of the nine labels is associated with 218 instruction steps (*i.e.,* the *labeled* dataset contains $218 \times 9 = 1962$ instruction steps).

Training: Due to the limited dataset size, we fine-tuned the model using the output classification layer of a 12-layer BERT model for uncased vocabulary, which has been used for generating contextual language embeddings [36]. We used Adam optimizer with the learning rate, ϵ, and batch size set to 2×10^{-5}, 10^{-8}, and 32, respectively, recommended by Devlin *et al.* [36]. 80%, 10% and 10% of the dataset are used for training, validation and testing, with the amount of steps for each key object balanced across the sets.

Model Performance: We demonstrated an overall 83.57% testing accuracy by considering the label generated by our rule-based approach as the ground truth (Fig. 6a). Additionally, we manually label the associated key objects on the testing dataset to limit the impact from any errors generated by our rule-based approach, which lead to a 82.13% overall accuracy (Fig. 6b).

Model Execution: The pre-trained model is used for predicting key objects from the segmented text of each step. To enhance the accuracy of the predicted key objects, we use prior knowledge provided while specifying the available key objects (Sec. 5.2). Specifically, the final assigned key object is the ML predicted key object with the highest confidence score that also belongs to the set of available key objects of the target environment.

6 CONSUMPTION PIPELINE

The consumption pipeline aims to spatialize each steps by anchoring them at the optimal position next to the key object. For example, consider how the instruction *"microwave on high for 30 seconds"* should be attached to a microwave. An ideal location would be at the front surface of the microwave door. A less idea location would be at the front of the input panel because the instruction might get in the way when the user tries to set the timer (see examples in Fig. 1c, Fig. 4i, and Fig. 20 in Appendix B.2).

6.1 Interaction Design

Our consumption pipeline provides dedicated interaction metaphors based on the preliminary findings (Sec. 3.2).

Navigating Between Individual Steps. Consumers can use hand menus to easily and rapidly switch between steps (Fig. 1b). We adopted the suggestions from **[DC1]** and the conceptual design of Fig. 3c that advocate the idea of delivering the right level of information only at the right time. Therefore, PaperToPlace only renders the current instruction step along with a task completion progress bar. When a new step is triggered, PaperToPlace first anchors the virtual label in front of the consumer, since a initial instruction step consuming is usually required before consumers proceeding on execute the associated steps (Fig. 3b).

Notations	Descriptions
W, H	Number of discretized cells along the width and height of the anchoring surface.
rot_s	Rotation, represented by quaternion, of the surface s.
du_s, dr_s, df_s	The up, right, and forward direction of the surface s.
p_{eye}	The midpoint of left and right eye in world coordinate.
df_{eye}	The forward direction of the gaze, averaged by left and right eye gaze.
$a = (r, c, s)$	Representation of an instruction step placement with respect to anchoring surface s, with index of r and c along width and height, where $r \in [0, W]$, $c \in [0, H]$.
p_a	The position of the instruction step in world coordinate.
$\theta_{x,y}$	The angle between vector x and y.

Table 1: Notations of the key parameters and functions.

Animating Spatial Guidance. To address **[DC3]**, we decided not using the persistent visual guidance (*e.g.,* virtual arrows) [48, 49] that might cause unnecessary visual disturbance. Instead, we use a animated flying effect where the virtual step could "fly" toward the key object after initial instruction step consuming. Such design leverage the fact that a motion effect could direct the consumers' attention, and could implicitly and rapidly offer visual guidance of the spatialized key object without causing overwhelming disturbance while consumers are executing the steps [45].

Placement of Instruction Steps. PaperToPlace places and anchors the instruction step on one of the anchoring surfaces of the key objects while not occluding the important region. This design emphasizes **[DC2]** suggesting the connections between instruction and real-world contexts, and could bring convenience while the consumers are attempting to refer back to the instruction step repeatedly while completing the step (Fig. 3c). If the consumer dislikes the label position, they can request a new position update on-demand using a mid-air pinch gesture. We also allow the consumers to use pinch-and-drag gestures to manually move the step to their preferred place (Fig. 1d). Such feedback action will in turn help on future decisions while placing instruction steps.

6.2 Problem Formulation

We formulated the process of optimally placing the instruction step as an optimization problem, where we used the tracked hands and gaze, as well as the anchoring surfaces defined by the spatial profile to search the optimal placement for each step. Table 1 summarizes the notations of key parameters.

Representations of an Instruction Step Placement. We first discretized each anchoring surface into $W \times H$ virtual cells where each cell has a dimension of $3cm \times 3cm$. We assumed that the center of each virtual step should be aligned with the center of the cell on the anchoring surface. We used $a = (r, c, s)$ to indicate the placement of a step, where r and c indicate the index of cell along the width and height of the surface s. The world position of the attempted step placement is $p_a = p_{topLeft} + 0.03 \cdot dr_s \cdot r - 0.03 \cdot du_s \cdot c$, where $p_{topLeft}$ is the position of the top left vertex of surface s.

Representations of the Step Label Rotation. Reading angles is a critical factor for consuming document [55]. It is therefore important to determine the rotation of the step, such that the text

Algorithm 1 Computing the rotation of instruction step.

1: **function** GetRotation(rot_s, p_a, p_{eye})
2:      $dir \leftarrow p_a - p_{eye}$ {Approximate potential looking direction.}
3:      $\alpha_{up} \leftarrow Angle(du_s, dir)$
4:      $\alpha_{right} \leftarrow Angle(dr_s, dir)$
5:      **if** s is a horizontal anchoring surface **then**
6:          **return** $rot_s * Quaternion.Euler(90° - \alpha_{up}, 0, 90° - \alpha_{right})$
7:      **else** {s is a vertical anchoring surface.}
8:          **return** $rot_s * Quaternion.Euler(90° - \alpha_{up}, \alpha_{right} - 90°, 0)$
9:      **end if**
10: **end function**
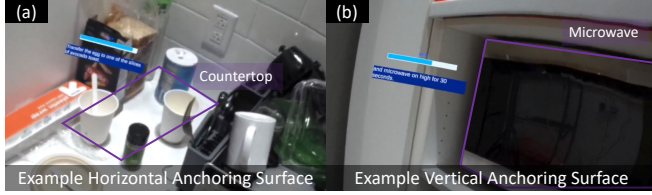
Figure 7: First-person view of rotating instruction step for (a) horizontal anchoring surface (*e.g.,* countertop) and (b) vertical anchoring surface (*e.g.,* microwave).

is always perpendicular to user's looking direction (*i.e.,* the virtual texts should be delivered facing toward user's eye). To address this, we rotated the anchoring surfaces by using the potential looking direction ($p_a - p_{eye}$). Algo. 1 shows how we compute the rotation of the step for horizontally (*e.g.,* countertop) and vertically placed anchoring surfaces (*e.g.,* the front surface of fridge). Fig. 7 demonstrates two examples where the steps are appropriately rotated.

Importance Map. One goal while attempting to place an instruction step onto associated anchoring surface(s) is to find the optimal placement that will not occlude consumer's interactions with the key object. While Lang *et al.* [51] presumed that the centroid of a key object is the most important area and should not be occluded by virtual MR agents, such assumption is invalid in our problem as the real-world interactions are highly dynamic. For example, while interacting with microwave, the critical areas might be the region on top of keypad when inputting cooking time, or the center areas when the user is checking whether the food is cooked. We used **importance map** (*imap*) to represent the importance of each possible cell on anchoring surface, where a larger *imap* value implies a higher probability that the corresponding position being interacted, and therefore should not be occluded by the virtual step. The values of *imap* are determined using near real-time data provided by the MR headset. By leveraging the pre-created spatial profile, we used the tracked gaze and hands to infer the importance of each possible position on anchoring surface(s). Intuitively, the areas near the hands, which usually imply the regions that are interacted by the consumers, might be more important and should not be occluded by the steps. Sec. 6.3 describes the computations of *imap*.

6.3 Importance Map on Anchoring Surfaces

Approximate the Importance of Each Frame. The contextual data from each **frame** refers to the tracked gaze and hand joints, at a specific time instant. Inferring *imap* from contextual data on single frame is less reliable, as real-world activities are highly dynamic.

Algorithm 2 Approximating the importance of individual frame.

1: **function** GetFrameWeights($v_{left}[\ ]$, $v_{right}[\ ]$, $t[\ ]$)
2:      $N \leftarrow Length(t)$ {N indicates the frame size.}
3:      $w, w_{speed}, w_{time} \leftarrow [0] * N, [0] * N, [0] * N$
4:      **for** $i \leftarrow 1$ to N **do**
5:          $w_{speed}[i] \leftarrow (|v_{left}[i]| + |v_{right}[i]|)/2$
6:          $w_{time}[i] \leftarrow (t[i] - t[1])/(t[N] - t[1])$
7:      **end for**
8:      $w_{speed} \leftarrow MinMaxNormalize(w_{speed})$ {Normalize to 0 to 1.}
9:      $w_{time} \leftarrow MinMaxNormalize(w_{time})$ {Normalize to 0 to 1.}
10:     $w \leftarrow Normalize((1 - w_{speed})w_{time})$ {$\Sigma_1^N(w[i])$ should be 1.}
11:   **return** w
12: **end function**

Yet, contextual data collected from a set of frames are practically not equally important. Therefore, while aggregating contextual data across a set of frames, it is important to consider the relative importance of individual frames. PaperToPlace uses the tracked eye behaviors to approximate the relative importance of contextual data of each frame. The *first* intuition is based on the instantaneous angular speed of gaze, where a slower angular speed of gaze implies a higher importance. For example, while attempting to input time when using microwave, the consumer might be *fixing* at the keypad, during which the contextual data could offer meaningful clues for approximating *imap*. Whereas, the consumer might rapidly *saccade* around the environment while finding ingredients, during which the contextual data might be less meaningful. While [56, 57] attempted to design closed-form solution for classifying saccade and fixation using the speed of gaze, such eye behaviours usually varies across users and tasks. Our *second* intuition is based on the observation that the contextual data from a more recent frame might be more useful to indicate the interactions in the subsequent task episode. Algo. 2 shows the computation of the **frame weight** (w) that is used to quantify the relative importance of contextual data at each frame. Experimentally, we set $N = 90$, which is approximately 1 second of past contextual data. We approximated w based on instantaneous angular speed of left and right gaze (v_{left}, v_{right}), and the timestamps (t) of each frames. Remarkably, a slower eye moves (*i.e.,* smaller v_{left} and v_{right}) and a more recent timestamp (*i.e.,* larger t) would lead to a more important frame weight. The $MinMaxNormalize(\cdot)$ and $Normalize(\cdot)$ represent the *min-max* normalization (to [0, 1]), and the normalization process such that the summation of the list is 1.

Approximate the Importance Map from Tracked Hands. To compute the overall *imap* from a set of frames, we first approximated the *imap* from the contextual data collected at each frame. This could be realized by $GetMap(\cdot)$ (Algo. 3), which *FloodFill* the *ConvexHull* of the projected points of 15 tracked hand joints [10] on the associated anchoring surfaces. Fig. 8d shows an example of the *imap* from single frame, approximated by tracked hands in Fig. 8b.

Generate the Overall Importance Map. While we assumed that the areas that are not occluded by hands might be less important, the importance assigned to each cell on the anchoring surfaces might not be equally same. For example, while inputting time using keyboard of microwave, the further the step being placed from the
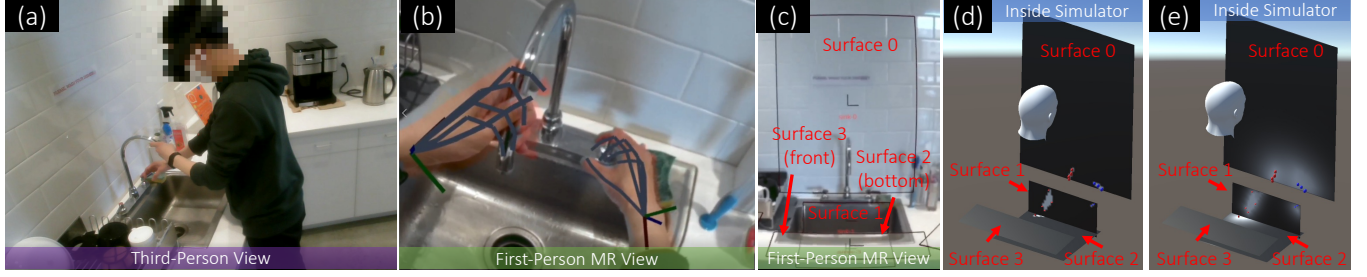
Figure 8: Examples of importance map. (a) Third-person view; (b) First-person view through MR; (c) First-person view of the key object (sink) containing four anchoring surfaces; (c) Importance map from single frame; (d) Overall importance map from a set of frames. The red and blue points in (d) and (e) indicate the projected key joints of the tracked left and right hand on the anchoring surfaces, respectively.

**Algorithm 3** Algorithms for computing importance map.

```
 1: function GETMAP(trackedJoints[], W, H)
 2:     hits, mask ← [], zeros(W, H)
 3:     for p_joint in trackedJoints do
 4:         hit ← RayCast(start : p_eye, direction : p_joint − p_eye)
 5:         if hit ≠ null then
 6:             w ← ToWidthIndex(hit.x)
 7:             h ← ToHeightIndex(hit.y)
 8:             mask[w, c] ← 1
 9:             hits.add([w, c])
10:         end if
11:     end for
12:     p_hull ← ConvexHull(hits)
13:     mask ← FloodFill(mask, p_hull)
14:     return mask
15: end function
16:
17: function GETOVERALLMAP(frames, W, H)
18:     map ← zeros(W, H)
19:     for i ← 1 to N do
20:         imap_left ← GetMap(frames[i].left.joints, W, H)
21:         imap_right ← GetMap(frames[i].right.joints, W, H)
22:         map ← map + w[i] * (Soft(imap_left) + Soft(imap_right))
23:     end for
24:     return MinMaxNormalize(map)  {Normalize to 0 to 1.}
25: end function
```

---

areas occluded by hands could lead to lower probability that the important areas being occluded by hands. Therefore, $imap$ should be expected to model *how important of a specific pixel on the anchoring surface*, instead of *whether the particular pixel is important*. Inspired by Lang *et al.* [51], we used $Soft(\cdot)$ to approximate the importance at each possible placement on the anchoring surface(s) generated by left and right hand respectively (Algo. 3). Eqn. 1 describes this process, where $e_{min}^{r,c}$ indicates the L2-distance from placement (r, c) to the closest placement(s) where the computed importance from individual frame from $GetMap(\cdot)$ (Algo. 3) is 1.

$$imap(r, c) = 1 - \frac{e_{min}^{r,c}}{max_{r',c'} e_{min}^{r',c'}} \quad (1)$$

The overall *imap* is finally computed by aggregating the *imap* generated by left ($imap_{left}$) and right ($imap_{right}$) hands on each frame using previous computed weight (w) in Algo. 2. This process could be demonstrated in $GetOverallMap(\cdot)$ (Algo. 3). Notably, if
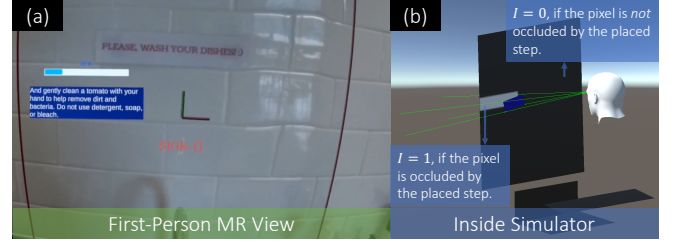
Figure 9: Example of occlusion map. (a) An example placement of the instruction step anchored next to the sink; (b) Visualization of the generated occlusion map, see Fig. 8c for the corresponding real-world scene.

there is no area being occluded by hands, we set $imap = O_{W,H}$. Fig. 8e shows an example overall *imap*.

6.4 Constraints and Costs

To solve the optimization problem for placing the instruction step on the anchoring surface(s), we need to model the constraints such that *the placed step will minimally occlude the user's view and will not be too far from the user's focused attention.*

Total Cost. We designed the overall cost ($C_{total}(a)$, Eqn. 2) as a weighted sum of the visibility cost ($C_V(a)$), the readability cost ($C_R(a)$), hand angle cost ($C_{HA}(a)$), and preference cost ($C_P(a)$). λ_V, λ_R, λ_{HA}, and λ_P are the weights associated with each of designed cost. Experimentally, we set them to 0.24, 0.24, 0.24 and 0.28, with slight emphasis on consumers' preference. We provide rationales of the design of each costs.

$$C_{total}(a) = \lambda_V C_V(a) + \lambda_R C_R(a) + \lambda_{HA} C_{HA}(a) + \lambda_P C_P(a) \quad (2)$$

(i) Visibility Cost. C_V aims to to measure how much key areas of the anchoring surfaces are occluded by a step placement a, and to penalize the situation while the step occluding the important areas. Eqn. 3 defines C_V, where I indicates **occlusion map** and $imap$ indicates the relative importance of each discretized cells on the anchoring surfaces computed by Algo. 3.

$$C_V(a) = \frac{[\sum_{r,c} I(r, c) \cdot imap(r, c)]^2}{||imap(r, c)||_2 \cdot \sum_{r',c'} I(r', c')} \quad (3)$$

Notably, the $I(r, c)$ is assigned as 1 when the pixel (r, c) is occluded by the step from the center of both eye (Fig. 9). Fig. 9b shows

an example occlusion map when the instruction step is placed next to the wall behind the sink. We finally normalized $C_V(a)$ to be independent of dimensions of anchoring surface(s).

(ii) Readability Cost. We penalized the solution when the delivered step is too far from the user's attention. To model this constraints, we used the eye tracking results and measure the $L2$-distance from the placed step to the weighted average of looking direction of both eye (df_{eye}). Eqn. 4 defines $C_R(a)$, where d_{max} represents the maximum $L2$-distance between two arbitrary solutions on the anchoring surfaces, which is computed by the maximum distance of the convex hull consisting of all vertices of the anchoring surfaces. Additionally, $C_R(a)$ need to enforce the instruction step is placed within binocular vision (*i.e.,* approximately $\pm 60°$) [58], to minimize the needs of moving head in order to read the instructions. We used a coefficient k to penalize the cost function, where k is set to 1 if $\theta_{p_a - p_{eye}, df_{eye}} < 60°$, otherwise we set $k = \theta_{p_a - p_{eye}, df_{eye}}$.

$$C_R(a) = \frac{k \cdot ||(p_a - p_{eye}) \cdot df_{eye} \cdot df_{eye} - (p_a - p_{eye})||_2}{d_{max}} \quad (4)$$

(iii) Hand Angle Cost. We modeled the observation that the instruction documents are usually held by and placed in front of the consumers' hands (Fig. 2). We first computed the angle between forward direction of the hand and the direction vector pointing from hand to the attempted solution ($\theta_{df_{hand}, p_a - p_{hand}}$), for left and right hand respectively (noted as $\theta_{leftHand}(i)$, $\theta_{rightHand}(i)$ computed from frame i). We then formulated the overall hand angle cost, by aggregating the angle cost generated by each frame (Eqn. 5).

$$C_{HA}(a) = \sum_i \frac{w[i] \cdot [\theta_{leftHand}(i) + \theta_{rightHand}(i)]}{360°} \quad (5)$$

(iv) Preference Cost. While PaperToPlace could determine the optimal step placement based on near-real-time context data, we retained the flexibility for users to specify their preferred placements. Eqn. 6 defines $C_P(a)$, where p_{pref,S_i} refer to the preferred step placement in the world coordinates for step S_i. We used $p_{pref,S_i} = null$ to indicate that the user has not manually fix the step placement. Similar to C_R, we used d_{max} to normalize preference cost.

$$C_P(a) = \begin{cases} \frac{||p_a - p_{pref,S_i}||_2}{d_{max}} & p_{pref,S_i} \neq null \\ 0 & p_{pref,S_i} = null \end{cases} \quad (6)$$

6.5 Optimizations

We aimed to minimize C_{total} by searching optimal placement for a specific step $\hat{a}$ (*i.e.,* $\hat{a} = \text{argmin}_a C_{total}(a)$). Finding optimal placement by computing C_{total} for each possibilities is impractical due to unnecessary latency and computational overhead. We instead used simulated annealing to approximate the global optimal [50].

Make a Move by Choosing the Neighbour Solution. We chose the neighbour solution $a^{i+1} = (r^i + \delta_r{}^i, c^i + \delta_c{}^i)$ by making a move of current solution $a^i = (r^i, c^i)$ at iteration i. Eqn. 7 describes our approach to determine the change over width ($\delta_r{}^i$) and height ($\delta_c{}^i$) of the anchoring surface at iteration i that yields smallest C_{total}, where $(\delta_r{}^i, \delta_c{}^i) \in \{(\delta_r, \delta_c) | \delta_r, \delta_c \in \{\pm 1, 0\} \wedge \delta_r \delta_c \neq 0\}$.

$$\delta_r, \delta_c = argmin_{\delta_{r'}, \delta_{r'}} C_{total}(a^{i+1} = (r^i + \delta'_r, c^i + \delta'_r)) \quad (7)$$
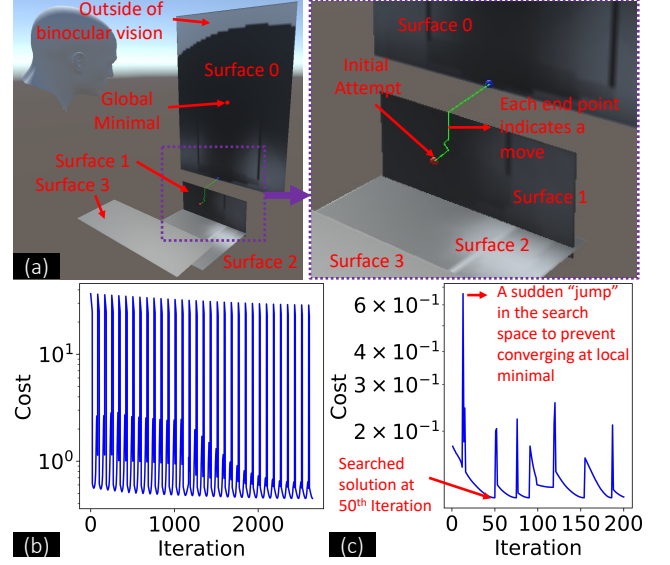
Figure 10: (a) Example placement attempts (the green trace). The darker color of the anchoring surface indicates a lower C_{total}; (b - d) Example optimized cost over each iteration using greedy algorithm (b) and simulated annealing approach (c). To increase readability, *log* scale is applied for y-axis. We only showed the traces before current cost reaching global minimal.

Notably, to ensure the sampled neighbouring solution is on the target anchoring surface, Eqn. 8 should be satisfied.

$$0 \leq r^{i+1} \leq W \quad 0 \leq c^{i+1} \leq H \quad r^{i+1}, c^{i+1} \in \mathbb{N} \quad (8)$$

Although such method performs well while making moves *within* single anchoring surface, the attempted solution will not be made *across* the surfaces. To address this, we specify that the placement on the neighbour anchoring surface, which is closest to the current placement attempt, would be chosen upon Eqn. 8 being violated. Fig. 10a demonstrates an example of how the neighbour solution is chosen while placement moving across the surfaces. To prevent converging at local minimal, we choose a random move in the global search space if $C_{total}(a^{i+1}) > C_{total}(a^i)$ [50].

Metropolis-Hastings Sampling and Simulated Annealing. We first selected a random placement on one of anchoring surface(s) randomly and used the Metropolis-Hastings algorithm [61] to sample the subsequent attempt. The probability for accepting the new attempt a^{i+1} is determined by Metropolis criteria. Eqn. 9 defines the the transition kernel of the Markov chain, where $T(i)$ indicates the temperature that will decay over the iteration.

$$p(a^{i+1}|a^i; i) = min\{1, exp(\frac{C_{total}(a^i) - C_{total}(a^{i+1})}{T(i)})\} \quad (9)$$

Notably, we used the empirical definition of $T(i)$, where $T(i) = \frac{T_1}{i+1}$ [50, 51]. Experimentally, we set $T_1 = 100$, and the number of iterations $i_{max} = 200$.

Comparisons of Optimization Performance. Fig. 10 demonstrates an example optimization result while attempting to place a step in front of sink that consist of four anchoring surfaces (Fig. 8c).
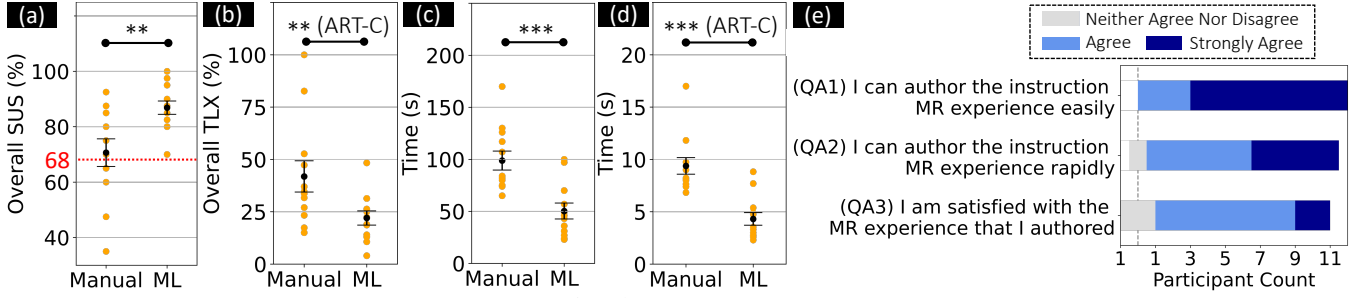
Figure 11: Results of authoring pipeline evaluations. (a - d) The overall SUS, weighted TLX scores, TCT of extracting document profiles, and the average task completion time for deciding each instruction step with segmented instruction step; (d) Survey results of how participants assessed the overall authoring pipeline and authored MR experience.

To better demonstrate the merits, we used greedy approach that tries each possibilities (Fig. 10b) and simulated annealing approach (Fig. 10c). Fig. 10a visualizes the C_{total} at each pixel on the anchoring surfaces with greedy approach, where darker area indicating a lower C_{total}. We showed that the greedy approach and simulated annealing approach need to make 2650 and 49 attempts respectively before finding the global minimum.

7 IMPLEMENTATIONS

We implemented the authoring pipeline on an iPad (9th generation) and the consumption pipeline on the Quest Pro [11] due to its colored passthrough, as well as eye [8] and hand tracking [10, 31] capabilities. We also implemented an optimization server using Unity 2021.3.9 [2] on a separate machine, and a Flask server for managing document (Sec. 5) and spatial profiles (Sec. 6).

8 USER STUDIES

Two *within-subject* studies were designed to evaluate PaperToPlace. 12 participants (PA1 - PA12) were recruited for evaluating the authoring pipeline and another 12 participants (PC1 - PC12) were recruited for evaluating consumption pipeline. During the evaluation, participants either authored or consumed the MR experiences for three cooking tasks (T1 - T3) that could be easily conducted in a typical office kitchen. Each study consists of two sessions where participants need to complete the designated tasks that involves with four key objects: *microwave*, *fridge*, *sink*, and *countertop* (Fig. 15). We used T1 as the training task, through which participants could get familiar with the designed interfaces. T2 and T3 were used for formal evaluations, each of which could be completed within 10 ∼ 15 min. Appendix A provides the details of instructions.

8.1 User Study 1: Authoring Pipeline

The first study aims to understand how well our authoring pipeline can support authors in easily and rapidly creating an MR instruction experience with manual or ML-assisted mode (Sec. 5.3). We aim to tackle three Research Questions (RQs):

- **(RQ1)** How the proposed authoring pipeline usable?
- How the ML could support **(RQ2)** a *faster* and **(RQ3)** an *easier* authoring experience?

Participants and Procedures. PA1 - PA12 (age, $M = 25.33, SD = 2.81$, *incl.* eight males and four females) were recruited. Six participants disagreed that they are experts of the designated tasks, with

the remaining participants held a neutral opinion. After training participants to use both of interfaces with T1, participants then completed one task for each interface condition (manual and ML-assisted). The tasks and interface conditions were counterbalanced across participants (Fig. 15 in Appendix B). Participants were finally instructed to briefly test the MR experience that was authored, to see if they were satisfied with the authoring outcomes inside MR. After each session, participants were asked to rate how strongly they agree with three prompts, shown in Fig. 11e, in a 5-point Likert scale. Participant were then invited to fill out the NASA TLX [46], followed by System Usability Scale (SUS) [26], as approximations of perceived workload and level of system usability. A semi-structured interview was conducted to understand participants' responses. The study on average took 37.87 min ($SD = 5.45$ min).

Measures. We analyzed the differences in the overall SUS, weighted TLX scores, and Task Completion Time (TCT) while extracting document profiles using both the manual and ML-supported modes. Shapiro-Wilk test [65] was used to check the normality of data in each catalogue. Repeated Measure Analysis of variance (RM-ANOVA) [43] with Tukey's HSD [19] ($\alpha = 0.05$) were used for analyzing statistical significance and *post-hoc* comparisons. Upon a failure of normality check, the Aligned Rank Transform (ART) [70], followed by ART contrast test (ART-C) [38] were used as the non-parametric approach for statistical significance analysis and *post-hoc* test[2]. The partial eta square (η_p^2) was used to evaluate the effect size for ART and RM-ANOVA, with .01, .06 and .14 indicating the thresholds for small, medium and large effect size [35]. We used thematic analysis [25], and deductive and inductive coding [39] to analyze qualitative data, to understand participants' experience.

Results and Discussions

Overall, most participants found the authoring pipeline easy to use (RQ1), and agreed that the ML-assisted mode could help the authoring experience faster (RQ2) and easier (RQ3).

Is the Proposed Authoring Pipeline Usable (RQ1)? Both manual ($M = 70.61, SD = 17.23$) and ML-assisted ($M = 86.88, SD = 8.33$) modes demonstrate good system usability, where an overall SUS score ≥ 68 is interpreted as "good" [62, 63]. All participants agreed that both pipelines are easy to use (Fig. 11a). For example: *"they are easy to use! [... and] integrated very good!"* (PA1) and *"it has a really good workflow"* (PA9). While 11 participant believed that they could

---
[2]Notations for indicating the *post-hoc* test results shown in Fig. 11, 14, and 13: ∗ ($p < .05$), ∗∗ ($p < .01$), ∗∗∗ ($p < .001$).
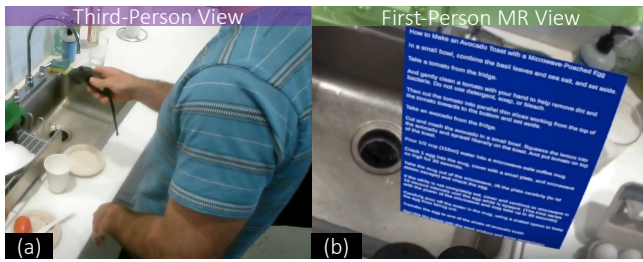
Figure 12: Baseline scene. Third-person (a) and first-person view through MR (b).

author such MR experience with both interfaces rapidly (Fig. 11e), PA10 held a neutral opinion, as she expected a fully automated system to extract the key objects associated with each instruction step. Overall, 10 participants were satisfied with the MR experience they authored (Fig. 11e). Examples include: *"a cool way to transfer knowledge"* (PA2) and *"useful to see [the instruction] while cooking"* (PA12). Few participants suggested the reason(s) for being not fully satisfied with the authored MR experience. For example, *"I think I made some mistake when I author it. So it guided me to the wrong spot"* (PA5) and *"I have to do [the tasks with authored experience] by myself, to see what it is like for me to experience that first before having a novice do it so"* (PA12). These implied the lack of ways to enable the authors to revise the authored experience inside MR iteratively.

Can ML Support a Faster Authoring Experience (RQ2)? The RM-ANOVA showed a reduced overall TCT ($F_{1,22} = 16.66$, $p < .001$, $\eta_p^2 = .43$, Fig. 11c) and average TCT (ART: $F_{1,22} = 35.98$, $p < .001$, $\eta_p^2 = .77$, Fig. 11d) for authoring each step while using ML-assisted mode, versus manual approach. Most participants echo such observations. For example: *"it could help me saving time, because [the key objects have] already [been] filled out [...] it speeds up by maybe a half a second [for each instruction]"* (PA3) and *"it helps me to save a lot of time! And that makes it a lot more convenient!"* (PA4)

Can ML Support an Easier Authoring Experience (RQ3)? The RM-ANOVA demonstrated a higher SUS ($F_{1,22} = 8.65$, $p = .008$, $\eta_p^2 = .28$, Fig. 11a) and a lower TLX score (ART: $F_{1,22} = 17.24$, $p = .002$, $\eta_p^2 = .61$, Fig. 11b) of the ML-assisted mode, compared to the manual counterpart. Most participants appreciated the convenience and helpfulness brought by the predicted key objects. First, nearly all participants believed that the ML helped on reducing effort for tagging key objects. For example: *"ML brought less effort, I just need to check if the predicted key object is correct or not [...] Even if I still need to check it, I don't have to pay 100% of attention. I don't have to do all the thing. I just have to do part of the thing"* (PA3), Particularly, the features of real-time predicting the new key object while modifying a specific step were favored by some participants. For example: *"when I saw one of the step to be very long and [are associated with two key objects] [...] capable of predicting key object after being modified is obviously helpful! And also the opposite feature where you could just like combine two tasks, followed by generating predicted key object! [...] it gives more flexibility while segmenting the instruction step"* (PA10).

Second, some participants highlighted the helpfulness for the mental thought process for ML-assisted mode. *"I was able to create*
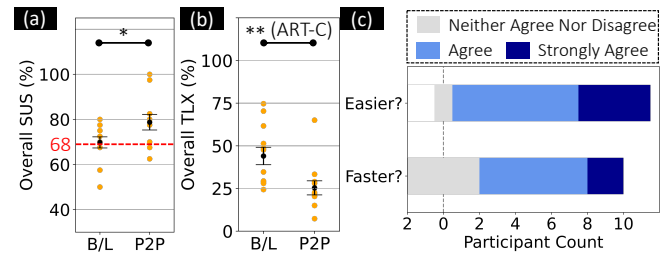
Figure 13: Consumption pipeline evaluation results. (a - b) Overall SUS and TLX score; (c) Survey results of how participants considered the overall consumption experience of PaperToPlace faster and easier, versus baseline. "B/L" and "P2P" indicate baseline and PaperToPlace condition.

a mental map of how I will spatially move across at different instances [by looking at the predicted key objects]" (PA4) and *"it could help me make a decision"* (PA3). Particularly, PA5 appraised the feature of predicting key object in real-time while revise the instruction: *"[while adding or editing the steps based on existing paper instruction], the real-time predicted location could help create a more clear instruction step. For example, if I type 'heat the water', and the predicted location is oven for somehow, then I might just type 'heat the water in the microwave' to make the instruction more clear"* However, PA10 held an opposite opinion: *"I just read the instruction step and then check if this assigned [key object] was countertop or not, and changed it to countertop rather than going and checking the other options. [...] but I did not read [the predicted key object] first."*

Finally, few participants suggested the merits of using color scale to visualize the confidence of predicted key object. For example: *"color could be helpful for conveying uncertainty"* (PA9) and *"color confidence is important to me. If it's red, I would be more aware of checking whether this is actually correct or not"* (PA12).

8.2 User Study 2: Consumption Pipeline

The second study aims to evaluate the consumption pipeline and attempts to address: *"how the PaperToPlace could help the consumers to complete the designated activities faster and easier?"*

Participants and Procedures. PC1 - PC12 (age, $M = 27.83$, $SD = 6.55$, *incl.* six males and six females) were recruited. All participants were *not* familiar with the designated tasks. We also built a baseline experience where the consumers could read existing monolithic instruction document inside MR (Fig. 12). Instead of asking participants to read a paper document, rendering a virtual monolithic document attached to the touch controller could minimize the impacts of confounding factors caused by uncomfort of the headset. We designed two sessions (Fig. 15) with counterbalancing being considered to minimize the impacts of prior learning experience and task familiarity. Before each session, T1 was used to help participants learn and familiarize with the system. During the session, participants were instructed to complete T2/T3 using the baseline and PaperToPlace. Participants were invited to fill out NASA TLX [46] and SUS [26] at the end of each session, followed by a semi-structured interview.

Measures. To understand the performance of context switching, we defined each **episode** as *the interval between the time when participants stopped a task to seek instructions and when they returned*
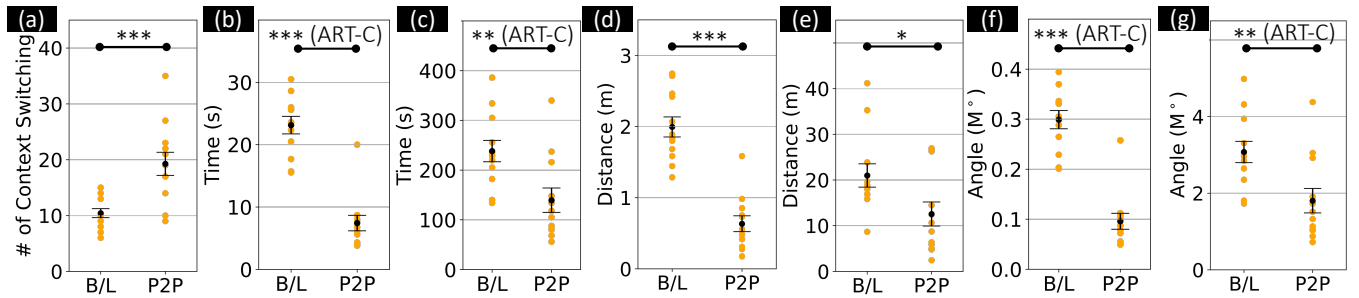
Figure 14: Results of context switching evaluations. (a) The total number of context switching while using the monolithic document and PaperToPlace. The average time (b), d_{head} (d), and θ_{head} (f) during each episode; The total time (c), l_{head} (e), and θ_{head} (g) during all episode while completing the task. "B/L" and "P2P" indicate the baseline and PaperToPlace conditions.

to the task. During each episode, we analyzed the **(i)** time; **(ii)** the distance of path of head movement (d_{head}); **(iii)** the angular changes of the forward direction of the head (θ_{head}). The data lies outside of such intervals are out of our scope, as the performance of real-world activities could be affected by participants' prior cooking experience. Same approaches in Sec. 8.1 were used to analyze the questionnaire responses and participant's qualitative feedback. The study on average took 59.40 min (SD = 5.80 min).

Results and Discussions. Overall, most participants believed that the PaperToPlace could help the consumers to complete the designated tasks faster and easier (Fig. 13c). Quantitatively, we demonstrated a higher overall SUS score ($F_{1,22}$ = 4.44, p = .046, η_p^2 = .17, Fig. 13a) and a lower perceived workloads (ART: $F_{1,22}$ = 18.52, p = 0.001, η_p^2 = .63, Fig. 13b). Based on participants' feedback, we now discuss how PaperToPlace could help participants complete the designated tasks faster and easier.

Context Awareness Reduces the Effort of Context-Switching. PaperToPlace reduces the average time (ART: $F_{1,22}$ = 49.18, p < .001 η_p^2 = .82, Fig. 14b), d_{head} ($F_{1,22}$ = 57.29, p < .001 η_p^2 = .72, Fig. 14d) and θ_{head} (ART: $F_{1,22}$ = 51.19, p < .001 η_p^2 = .82, Fig. 14f) on each episode. While PaperToPlace on average leads to frequent document readings ($F_{1,22}$ = 15.77, p < .001, η_p^2 = .42, Fig. 14a), the accumulated time (ART: $F_{1,22}$ = 9.20, p = .020 η_p^2 = .42, Fig. 14c), d_{head} ($F_{1,22}$ = 5.30, p = .030 η_p^2 = .19, Fig. 14e) and θ_{head} (ART: $F_{1,22}$ = 7.48, p = .019 η_p^2 = .40, Fig. 14g) of all episodes are reduced.

First, participants suggested the convenience for referring back to the instructions repeatedly with PaperToPlace. For example: *"[with baseline], I need to check it back and forth every time while trying to grab food from the fridge [...] [PaperToPlace] gives me feeling like [the instruction] is just on my side. It's like always on my side, like right beside my head"* (PC9) and *"because [the step] would always be right there with just a little bit of information I need, I think it'd be very useful"* (PC11). Most participants explicitly highlighted the benefits of finding optimal placements, without causing occlusions to the key interaction areas. For example: *"it is useful to bring the instruction step to me by just a pinch."* (PC10), *"I like the position of instruction step!"* (PC4), and *"I think it is useful! And especially the function of where you pinch again, it will move to another location, so it can ensure [the step] will never block your sight"* (PC8). More examples of instruction step placements could be referred to Fig. 20e - h.

However, PC6, who unveiled her ADHD [34], suggested: *"I was distracted with [the PaperToPlace], because [when the step occasionally was not anchored on the optimal position] the information was here and I was there. It reminded me of the moments that I forgot what I was supposed to do, or what I have to do."*

Second, most participants acknowledged the helpfulness by establishing connections between instructions and key objects. For example: *"I like how it took me to the sink because this activity has to be near the sink. That's a very helpful on spatial understanding!"* (PC4), *"Although we know where the fridge is, having that is really convenient to just not give anything a thought and do things as per the instructions"* (PC10).

Finally, five participants also mentioned the merits of hands-free of PaperToPlace, compared to the baseline where the consumers need to hold the virtual document. For example *"[baseline] is more cumbersome, because I need to free one hand and make the hand very clean to make sure that the hand is clean to touch the controller"* (PC5).

Segmented Instruction Helps Findings the Relevant Information Easier. All participants suggested that the segmented instructions is helpful, *e.g., "[instructions] need to be as concise and as short as possible to be read at the same time. [PaperToPlace] did its job!"* (PC5). Most participants suggested the merits of reducing stress while translating instruction into real world activities. For example: *"I have more calmness [with PaperToPlace], because [with baseline] I was seeing everything all at once, and that was giving me the feeling I'm in a hurry."* (PC7), and *"looking at the entire document at once was so hard that I forgot where I have to keep following from start to finish to find where I was. But [PaperToPlace] gave me one by one instructions, which is super easy!"* (PC3)

9 LIMITATIONS AND FUTURE WORKS

We identify our limitations from four perspectives.

(1) Enabling an Iterative Authoring Process. We observed that during authoring, participants intended to segment the instruction steps by *only* considering whether only one key object is associated with the specific step, without synthesizing other factors (*e.g.,* the density of the information contained by single step while viewed inside MR). However, this cannot ensure a satisfied MR experience from the perspective of the consumers. Future work might investigate potential *iterative* authoring workflow that allows the authors to refine their authored document profile inside MR while piloting the created instructional MR experience.

(2) Transforming Richer Metadata into MR Experiences. We consider that the metadata of each instruction step *only* contains the text of the step and the key objects that the virtual step should be anchored on (Sec. 4.1). This might not be realistic for real-world instruction documents with heterogeneous kinds of metadata, such as the duration information, the caveats that usually requires the consumers' attention, and the notifications from the environmental sensors. For example: *"I would like to have warning text, like 'do not use detergent', maybe show up in a different color or something"* (PC4). Future research might investigate the richer metadata that need to be augmented inside the MR experience, and the methods to use existing language models to extract such metadata as well as transform them into spatialized and context-aware MR experience.

(3) Automatically Switching between Instruction Steps and Triggering the Position Update of the Instruction Label. We currently required consumers to explicitly click the virtual button to switch to the next instruction step, and to pinch to update the current position of the instruction step on demand (Sec. 6.1). While participants (*e.g.,* PC10) with some prior MR experience felt it is *"easy and useful"* to use the virtual hand menu and pinch gesture, others (*e.g.,* PC1) suggested the frustrations of occasional failures of pinch gesture detections and the virtual button clicking. Future work might consider designing a state machine, which could specify how to switch to the subsequent step *automatically* based on user's activities that might be inferred from face (*e.g.,* [28]), body (*e.g.,* [5]) and environmental (*e.g.,* [20, 24, 27]) sensor data.

(4) Supporting a Broader Range of Applications. While many participants believed *"cooking demonstrates [PaperToPlace] very well"* (PC4), we only evaluated on cooking instructions, due to the poor quality of passthrough capabilities of Quest Pro [11]; availability of dataset to fine tune language model for alternative instruction activities; and limited study resources. Future work might explore other activities with more powerful language model such as GPT and prompt engineering techniques being used for creating document profiles. Participants also emphasized the values of adaptive placements (*e.g.,* *"the adaptive placement of instruction is definitely useful for paper cutting! I don't want to cut my hand. And I wanted the instruction to be always besides my hand"* (PC9)) and reduced context switching (*e.g.,* *"in the gym, where I need some instructions to teach me how to use the equipment, such as how you hold the gears with good postures. [...] with [baseline], it is less efficient and [I have] to stop in between and read the instructions"* (PC2)) that might be transferred to other activities. Another direction is to investigate the support for finer grained tasks that might be involved with moving objects, leading to a dynamically changed spatial profile. For example, *"if you are doing PCB soldering, it might be hard to track that tiny component and to pinpoint the exact location on the PCB board. But if [PaperToPlace] can do that, it will be super helpful!"* (PC8). This requires high quality passthrough and capabilities to track real-time location of the electronic components which are considered as *non*-static key objects. Instead of using Quest Pro, future researchers might consider a more recent higher-end headset, *e.g.,* Vision Pro [18].

10 CONCLUSION

We present and evaluate PaperToPlace, comprising *an authoring pipeline*, which allows authors to rapidly transform existing paper instructions into a MR experience, and *a consumption pipeline*, which enables consumers to view spatialized instructions using a context-aware approach. Two within-subject studies with two different cohorts of 12 participants demonstrate the usability and effectiveness of the proposed authoring and consumption workflows.

ACKNOWLEDGMENTS

REFERENCES

[1] 2016. *Receipe: Microwave Mac 'n' Cheese in a Mug*. https://www.newideafood.com.au/microwave-mac-n-cheese-in-a-mug Accessed: 12-24-2022.

[2] 2021. *Unity 2021.3.9 Release*. https://unity.com/releases/editor/whats-new/2021.3.9 Accessed: 04-01-2023.

[3] 2022. *Adobe Scan Mobile Apps*. https://www.adobe.com/acrobat/mobile/scanner-app.html Accessed: 08-29-2022.

[4] 2022. *Basic Microwave Scrambled Eggs*. https://www.eggs.ca/recipes/basic-microwave-scrambled-eggs Accessed: 12-24-2022.

[5] 2022. *Body Tracking for Movement SDK for Unity*. https://developer.oculus.com/documentation/unity/move-body-tracking/ Accessed: 07-23-2023.

[6] 2022. *Bring Surroundings into Your App with RoomPlan*. https://developer.apple.com/augmented-reality/roomplan/ Accessed: 12-24-2022.

[7] 2022. *Cuisinart Instruction Booklet*. https://www.cuisinart.ca/on/demandware.static/-/Sites-ca-cuisinart-Library/en_CA/dwca81d85d/Instruction%20Booklet/DGB-800C_en.pdf Accessed: 12-24-2022.

[8] 2022. *Eye Tracking for Movement SDK for Unity*. https://developer.oculus.com/documentation/unity/move-eye-tracking/ Accessed: 12-24-2022.

[9] 2022. *Full Text Annotation of Google Vision OCR API*. https://cloud.google.com/vision/docs/fulltext-annotations Accessed: 12-24-2022.

[10] 2022. *Hand Tracking*. https://developer.oculus.com/documentation/unity/unity-handtracking/ Accessed: 12-24-2022.

[11] 2022. *Meta Quest Pro*. https://www.meta.com/quest/quest-pro/ Accessed: 12-24-2022.

[12] 2022. *Microsoft Dynamic 365 Remote Assist*. https://dynamics.microsoft.com/en-us/mixed-reality/guides/ Accessed: 07-04-2023.

[13] 2022. *OCR Provided by Google Vision API*. https://cloud.google.com/vision/docs/ocr Accessed: 12-24-2022.

[14] 2022. *Receipe: Chocolate Cake in a Mug*. https://www.foodnetwork.com/recipes/ree-drummond/chocolate-cake-in-a-mug-3158576 Accessed: 12-24-2022.

[15] 2022. *Receipe: Quick, Microwave-Poached Eggs on Avocado Toast*. https://blog.myfitnesspal.com/quick-microwave-poached-eggs-on-avocado-toast Accessed: 12-24-2022.

[16] 2022. *Spatial Anchors Overview*. https://developer.oculus.com/documentation/unity/unity-spatial-anchors-overview/ Accessed: 12-24-2022.

[17] 2022. *Tab – The Simple Bill Splitting Apps*. https://www.tabapp.co Accessed: 08-29-2022.

[18] 2023. *Apple Vision Pro*. https://www.apple.com/apple-vision-pro/ Accessed: 07-04-2023.

[19] Hervé Abdi and Lynne J. Williams. 2010. Tukey's Honestly Significant Difference (HSD) Test. *Encyclopedia of Research Design* 3, 1 (2010), 1–5.

[20] Yuvraj Agarwal, Christopher Harrison, Gierad Laput, Sudershan Boovaraghavan, Chen Chen, Abhijit Hota, Bo Robert Xiao, and Yang Zhang. 2019. Virtual Sensor System. https://patents.google.com/patent/US10436615B2 US Patent 10,436,615.

[21] John R. Anderson. 2013. *The Architecture of Cognition*. Psychology Press.

[22] Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. RecipeNLG: A Cooking Recipes Dataset for Semi-Structured Text Generation. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, 22–28. https://aclanthology.org/2020.inlg-1.4

[23] Stephen P. Bogdewic. 1992. Participant Observation. (1992).

[24] Sudershan Boovaraghavan, Chen Chen, Anurag Maravi, Mike Czapik, Yang Zhang, Chris Harrison, and Yuvraj Agarwal. 2023. Mites: Design and Deployment

of a General-Purpose Sensing Infrastructure for Buildings. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 2 (mar 2023), 32 pages. https://doi.org/10.1145/3580865

[25] Virginia Braun and Victoria Clarke. 2012. *Thematic Analysis.* American Psychological Association.

[26] John Brooke. 1996. SUS – A Quick and Dirty Usability Scale. *Usability Evaluation in Industry* 189, 194 (1996), 4–7.

[27] Chen Chen, Ke Sun, and Xinyu Zhang. 2020. CapTag: Toward Printable Ubiquitous Internet of Things: Poster Abstract. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems* (Virtual Event, Japan) *(SenSys '20).* Association for Computing Machinery, New York, NY, USA, 669–670. https://doi.org/10.1145/3384419.3430410

[28] Chen Chen, Ke Sun, and Xinyu Zhang. 2021. ExGSense: Toward Facial Gesture Sensing with a Sparse Near-Eye Sensor Array. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (Co-Located with CPS-IoT Week 2021)* (Nashville, TN, USA) *(IPSN '21).* Association for Computing Machinery, New York, NY, USA, 222–237. https://doi.org/10.1145/3412382.3458268

[29] Chen Chen, Matin Yarmand, Varun Singh, Michael V. Sherer, James D. Murphy, Yang Zhang, and Nadir Weibel. 2022. Exploring Needs and Design Opportunities for Virtual Reality-Based Contour Delineations of Medical Structures. In *Companion of the 2022 ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Sophia Antipolis, France) *(EICS '22 Companion).* Association for Computing Machinery, New York, NY, USA, 19–25. https://doi.org/10.1145/3531706.3536456

[30] Chen Chen, Matin Yarmand, Varun Singh, Michael V. Sherer, James D. Murphy, Yang Zhang, and Nadir Weibel. 2022. VRContour: Bringing Contour Delineations of Medical Structures Into Virtual Reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).* 64–73. https://doi.org/10.1109/ismar55827.2022.00020

[31] Chen Chen, Matin Yarmand, Zhuoqun Xu, Varun Singh, Yang Zhang, and Nadir Weibel. 2022. Investigating Input Modality and Task Geometry on Precision-first 3D Drawing in Virtual Reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).* 384–393. https://doi.org/10.1109/ISMAR55827.2022.00054

[32] 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* (Virtual Event, USA) *(UIST '21).* Association for Computing Machinery, New York, NY, USA, 282–297. https://doi.org/10.1145/3472749.3474750

[33] 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 *Designing Interactive Systems Conference 2021.* 234–249. https://doi.org/10.1145/3461778.3462126

[34] Mayo Clinic. 2023. *Adult Attention-Deficit/Hyperactivity Disorder (ADHD).* https://www.mayoclinic.org/diseases-conditions/adult-adhd

[35] H Cohen. 1988. Statistical Power Analysis for Behavioral Sciences. *Hillsdale, NJ: Lawrence Erlbaum Associates* (1988).

[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arxiv.1810.04805

[37] Elsa Eiriksdottir and Richard Catrambone. 2011. Procedural Instructions, Principles, and Examples: How to Structure Instructions for Procedural Tasks to Enhance Performance, Learning, and Transfer. *Human Factors* 53, 6 (2011), 749–770. https://doi.org/10.1177/0018720811419154 arXiv:https://doi.org/10.1177/0018720811419154 PMID: 22235535.

[38] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21).* Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[39] Satu Elo and Helvi Kyngäs. 2008. The qualitative content analysis process. *Journal of advanced nursing* 62, 1 (2008), 107–115.

[40] Gerhard Fischer. 2012. Context-Aware Systems: The 'right' Information, at the 'Right' Time, in the 'Right' Place, in the 'Right' Way, to the 'Right' Person. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island, Italy) *(AVI '12).* Association for Computing Machinery, New York, NY, USA, 287–294. https://doi.org/10.1145/2254556.2254611

[41] Ran Gal, Lior Shapira, Eyal Ofek, and Pushmeet Kohli. 2014. FLARE: Fast Layout for Augmented Reality Applications. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).* 207–212. https://doi.org/10.1109/ismar.2014.6948429

[42] Franck Ganier. 2004. Factors Affecting the Processing of Procedural Instructions: Implications for Document Design. *IEEE Transactions on Professional Communication* 47, 1 (2004), 15–26. https://doi.org/10.1109/tpc.2004.824289

[43] Ellen R. Girden. 1992. *ANOVA: Repeated Measures.* Number 84. Sage University Paper Series.

[44] MIS Guntur, W Setyaningrum, H Retnawati, et al. 2020. Can augmented reality improve problem-solving and spatial skill?. In *Journal of Physics: Conference Series,* Vol. 1581. IOP Publishing, 012063. https://doi.org/10.1088/1742-6596/1581/1/012063

[45] Aurora Harley. 2014. *Animation for Attention and Comprehension.* https://www.nngroup.com/articles/animation-usability

[46] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA–TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload,* Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[47] Robert Johansen. 1988. *GroupWare: Computer Support for Business Teams.* The Free Press, USA. https://doi.org/10.5555/542298

[48] Janet G Johnson, Danilo Gasques, Tommy Sharkey, Evan Schmitz, and Nadir Weibel. 2021. Do You Really Need to Know Where "That" Is? Enhancing Support for Referencing in Collaborative Mixed Reality Environments. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 514, 14 pages. https://doi.org/10.1145/3411764.3445246

[49] Janet G. Johnson, Danilo Gasques Rodrigues, Madhuri Gubbala, and Nadir Weibel. 2018. HoloCPR: Designing and Evaluating a Mixed Reality Interface for Time-Critical Emergencies. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare* (New York, NY, USA) *(PervasiveHealth '18).* Association for Computing Machinery, New York, NY, USA, 67–76. https://doi.org/10.1145/3240925.3240984

[50] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. 1983. Optimization by Simulated Annealing. *Science* 220, 4598 (1983), 671–680. https://doi.org/10.1126/science.220.4598.671

[51] Yining Lang, Wei Liang, and Lap-Fai Yu. 2019. Virtual Agent Positioning Driven by Scene Semantics in Mixed Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR).* 767–775. https://doi.org/10.1109/VR.2019.8798018

[52] Wei Liang, Xinzhe Yu, Rawan Alghofaili, Yining Lang, and Lap-Fai Yu. 2021. Scene-Aware Behavior Synthesis for Virtual Pets in Mixed Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 472, 12 pages. https://doi.org/10.1145/3411764.3445532

[53] 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* (New Orleans, LA, USA) *(UIST '19).* Association for Computing Machinery, New York, NY, USA, 147–160. https://doi.org/10.1145/3332165.3347945

[54] Jingjing Liu, Wei Liane, Bing Ning, and Ting Mao. 2021. Work Surface Arrangement Optimization Driven by Human Activity. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR).* 270–278. https://doi.org/10.1109/VR50410.2021.00049

[55] Meredith Ringel Morris, A.J. Bernheim Brush, and Brian R. Meyers. 2007. Reading Revisited: Evaluating the Usability of Digital Display Surfaces for Active Reading Tasks. In *Second Annual IEEE International Workshop on Horizontal Interactive Human-Computer Systems (TABLETOP'07).* 79–86. https://doi.org/10.1109/tabletop.2007.12

[56] Matthew S. Mould, David H. Foster, Kinjiro Amano, and John P. Oakley. 2012. A simple nonparametric method for classifying eye fixations. *Vision Research* 57 (2012), 18–25. https://doi.org/10.1016/j.visres.2011.12.006

[57] Marcus Nyström and Kenneth Holmqvist. 2010. An Adaptive Algorithm for Fixation, Saccade, and Glissade Detection in Eyetracking Data. *Behavior research methods* 42, 1 (2010), 188–204. https://doi.org/10.3758/brm.42.1.188

[58] John D Pettigrew. 1986. The evolution of binocular vision. *Visual neuroscience* (1986), 208–222.

[59] Jing Qian, Qi Sun, Curtis Wigington, Han L. Han, Tong Sun, Jennifer Healey, James Tompkin, and Jeff Huang. 2022. Dually Noted: Layout-Aware Annotations with Smartphone Augmented Reality. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 552, 15 pages. https://doi.org/10.1145/3491102.3502026

[60] 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* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 65, 18 pages. https://doi.org/10.1145/3491102.3517665

[61] Christian P Robert and George Casella. 1999. The Metropolis—Hastings algorithm. In *Monte Carlo statistical methods.* Springer, 231–283.

[62] Jeff Sauro. 2011. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices.* Measuring Usability LLC.

[63] Jeff Sauro and James R Lewis. 2016. *Quantifying the User Experience: Practical Statistics for User Research.* Morgan Kaufmann.

[64] Karen A. Schriver. 1997. *Dynamics in Document Design: Creating Text for Readers.* John Wiley & Sons, Inc.

[65] Samuel Sanford Shapiro and Martin Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 3-4 (dec 1965), 591–611. https://doi.org/10.1093/biomet/52.3-4.591
[66] Richard Skarbez, Missie Smith, and Mary C Whitton. 2021. Revisiting milgram and kishino's reality-virtuality continuum. *Frontiers in Virtual Reality* 2 (2021), 647997. https://doi.org/10.3389/frvir.2021.647997
[67] James P Spradley. 2016. *Participant Observation.* Waveland Press.
[68] 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.* Association for Computing Machinery, New York, NY, USA, 328–341. https://doi.org/10.1145/3379337.3415815
[69] Ann C Weller. 1986. *How to Write a Usable User Manual.* Vol. 74. Medical Library Association. 274 pages.
[70] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11).* Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963
[71] Hongfei Yu, Wei Liang, Shihao Song, Bing Ning, and Yixin Zhu. 2021. Interactive Context-Aware Furniture Recommendation using Mixed Reality. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW).* 450–451. https://doi.org/10.1109/vrw52623.2021.00107

Step Descriptions
● Spray microwave-safe container (*e.g.,* mug, ramekin, or egg cooker) with cooking spray or wipe lightly with vegetable oil.
● Whisk eggs, milk, salt and pepper in container (or whisk ingredients in another bowl and pour into microwave container). If using a mug or ramekin, cover with plastic wrap, pulling back small area for venting. If using an egg cooker, place lid on cooker base, lining up notches. Twist to secure.
● Microwave on Medium-High (70% power) for 90 seconds, stirring several times during cooking.
● Cover and let stand for 30 seconds to 1 minute before serving. Eggs will look slightly moist, but will finish cooking upon standing.

Table 2: Experimental cooking recipe for making basic microwave scrambled eggs (T1).

A EXPERIMENTAL TASKS

The selected tasks for final user study (Sec. 8) include:

- **(T1)** Microwave Scrambled Eggs [4];
- **(T2)** Quick Microwave-Poached Eggs on Avocado Toast [15];
- **(T3)** Instant Mac 'n' Cheese [1];

Table 2, Table 3 and Table 4 offers the supplementary material regarding the specific instruction steps of the experimental tasks T1, T2, and T3 respectively. Notably, T1 was used as the training tasks for participants to learn and get familiar with the interfaces (Table 2). Assuming the results from OCR is fully correct (*i.e.,* all texts of all instruction steps could be successfully extracted), the overall accuracies of the associated key objects predicted by our pre-trained language model for each experimental tasks are 50% (T1), 84.62% (T2), and 80.00% (T3). Notably, the overall accuracies for T2 and T3 are closed to our benchmark results while fine-tuning the BERT model in Sec. 5.3, which is 82.13%. This ensures the results yielded by the evaluation of authoring pipeline (Sec. 8.1) is generalizable to some extend. Although the overall accuracy of T1 is far lower than our benchmark results due to the relative short of instruction document, the instruction of T1 was only used for participants to familiarize themselves with the given interfaces (either on iPad or inside MR), and the data yielded by T1 was excluded from our evaluation results in Sec. 8.

B USER STUDY RESULTS

This section presents supplementary material for Sec. 8. Fig. 15 shows the specific tasks and interface conditions that were assigned while evaluating authoring and consumption pipeline. Notably, T1 was used for training purposes.

We also provide visualizations of survey responses for authoring pipeline evaluations (Sec. B.1) as well as consumption evaluations (Sec. B.2).

B.1 Evaluations of Authoring Pipeline

Fig. 16 demonstrates the NASA TLX responses of each perceived workloads from participants PA1 - PA12, while using manual and ML-assisted interfaces to extract document profile from designated paper instruction. Fig. 18 provides supplementary material of survey results of SUS questionnaires. For Fig. 16 and Fig. 18, we use

Step Descriptions

- In a small bowl, combine the basil leaves and sea salt, and set aside.
- Take a tomato from the fridge.
- And gently clean a tomato with your hand to help remove dirt and bacteria. Do not use detergent, soap, or bleach.
- Then cut the tomato into parallel thin slices working from the top of the tomato towards to the bottom and set aside.
- Take an avocado from the fridge.
- Cut and mash the avocado in a small bowl. Squeeze the lemon into the avocado and spread liberally on the toast. And put tomato on top of the toast.
- Pour 1/2 cup (118 ml) water into a microwave-safe coffee mug.
- Crack 1 egg into the mug, cover with a small plate, and microwave on high for 30 seconds.
- Take the mug out of the microwave, lift the plate carefully (to let steam escape) and check the egg.
- If the white is not completely set, cover and continue to microwave in 10-second intervals until the egg white is opaque. (The time varies with the power of the microwave and may take up to 60 seconds).
- Carefully pour off the water in the mug, using a slotted spoon to keep the egg from falling out.
- Transfer the egg to one of the slices of avocado toast.
- Sprinkle the toasts with the seed mixture and serve immediately.

Table 3: Experimental cooking recipe for making an avocado toast with a microwave-poached egg (T2).

Step Descriptions

- Find a mug that holds twice the volume of your dry pasta – the bigger, the better.
- Add the macaroni.
- Add some water.
- Cover with cling film and pierce 3 times.
- Stand the mug in a microwave-proof bowl to catch any spillages, and cook in the microwave on high for 2 minutes. The liquid will bubble up and over the sides, so tip any liquid from the bowl back into the mug (be careful as it will be very hot) and give it a good stir.
- Leave to stand for 1 minute.
- Repeat twice more or until the pasta is cooked (it may take longer depending on the pasta).
- Then remove from the microwave.
- Repeat twice more or until the pasta is cooked (it may take longer depending on the pasta).
- Stir through the butter, cheese and spinach or Marmite, if using.
- The heat from the pasta should melt the cheese and wilt the spinach, but if not, pop back in the microwave for 30 seconds.

Table 4: Experimental cooking recipe for making microwaved mac 'n' cheese (T3).
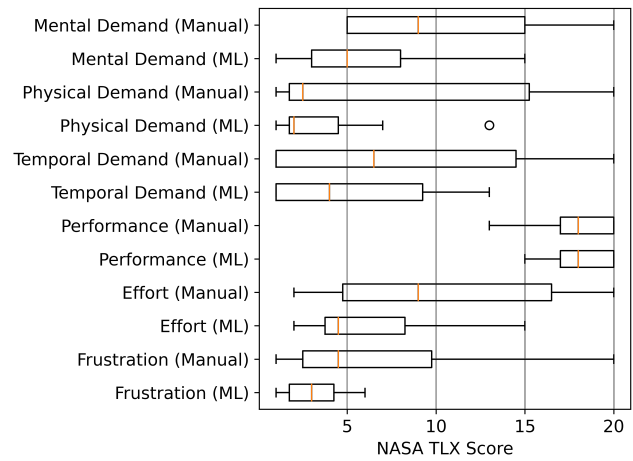
"Manual" and "ML" to indicate the interface condition while manual and ML supported approaches are used while extracting the associated key objects from each instruction steps, respectively.

## B.2 Evaluations of Consumption Pipeline

Fig. 17 demonstrates the NASA TLX responses of each perceived workloads from participants PC1 - PC12, while using baseline and PaperToPlace interfaces to perform the designated tasks. Fig. 19

Study 1: Authoring Pipeline			Study 2: Consuming Pipeline		
Participant	Session 1	Session 2	Participant	Session 1	Session 2
PA1 – PA3	Without ML Supported, T2	With ML Supported, T3	PC1 – PC3	Non-Context-Aware & Non-Spatialized, T2	Context-Aware & Spatialized, T3
PA4 – PA6	Without ML Supported, T3	With ML Supported, T2	PC4 – PC6	Non-Context-Aware & Non-Spatialized, T3	Context-Aware & Spatialized, T2
PA7 – PA9	With ML Supported, T2	Without ML Supported, T3	PC7 – PC9	Context-Aware & Spatialized, T2	Non-Context-Aware & Non-Spatialized, T3
PA10 – PA12	With ML Supported, T3	Without ML Supported, T2	PC10 – PC12	Context-Aware & Spatialized, T3	Non-Context-Aware & Non-Spatialized, T2

Figure 15: Study design for evaluating PaperToPlace. Each participant needs to conduct session 1 and session 2 in order. T2 and T3 were used for formal evaluation while T1 was used for training purposes.
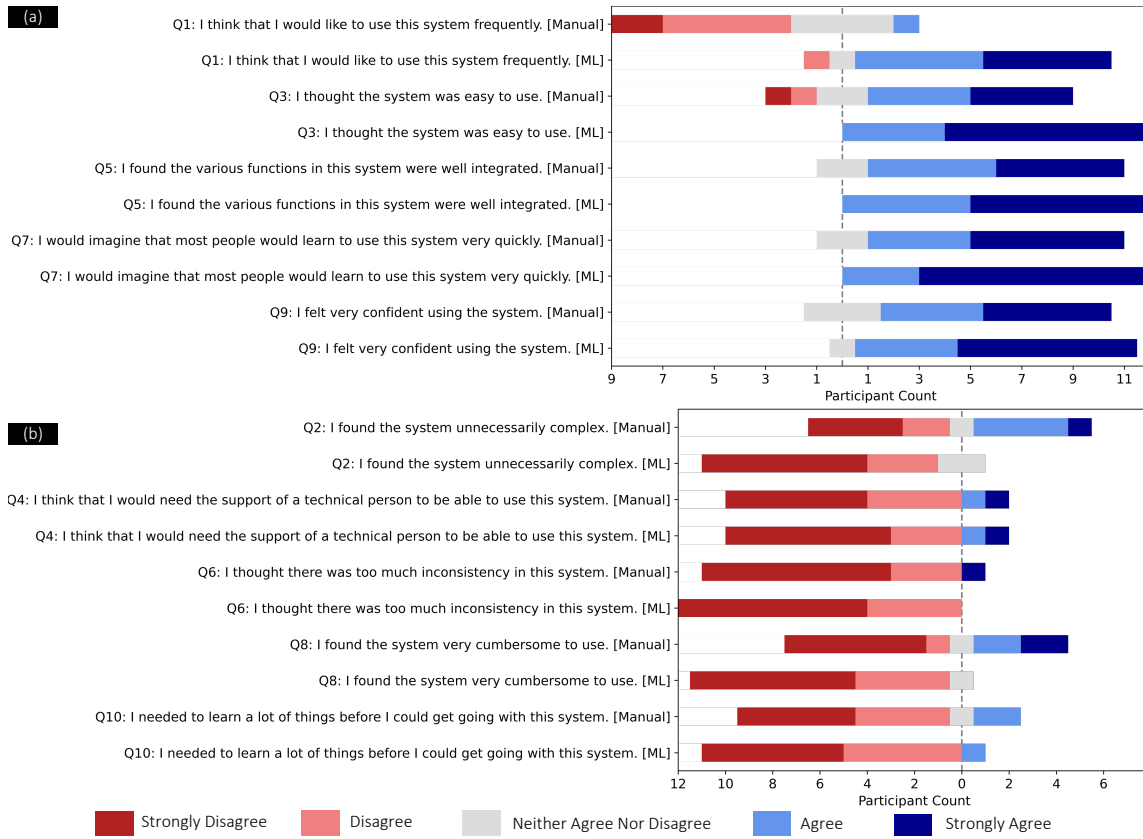

Figure 16: Survey results of NASA TLX questionnaires. We use "Manual" and "ML" to indicate the interface condition while manual and ML supported approaches are used for extracting associated key objects from each instruction steps, respectively
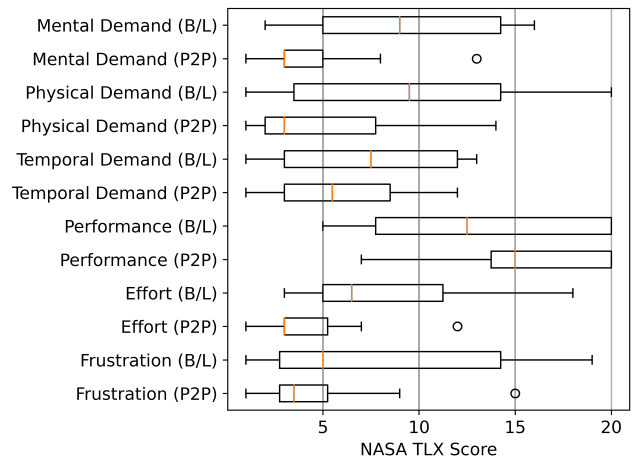
Figure 17: Survey results of NASA TLX questionnaires. We use "B/L" to refer to the baseline interface, and "P2P" to indicate PaperToPlace, which delivers spatialized and context-aware instruction step.

provides supplementary material of survey results of SUS questionnaires. To be consistent with the remaining of this paper, we

Figure 18: Survey results of SUS questionnaires of authoring pipeline evaluations. We use "Manual" and "ML" to indicate the interface condition while manual and ML support approaches are used for extracting associated key objects from each instruction steps, respectively. To increase readability, we cluster the survey results of positive statements (Q1, Q3, Q5, Q7, Q9) into subplot (a), where a higher level of agreement indicates a better user experience. The survey results of negative statements (Q2, Q4, Q6, Q8, Q10) are clustered into subplot (b), where a lower level of agreement indicates a better user experience.

use "B/L" to refer to the baseline interface, and "P2P" to indicate PaperToPlace, which delivers spatialized and context-aware instruction step. Finally, Fig. 20e - h showed examples of how the virtual instructions steps would be anchored on the key objects with the consumption pipeline of PaperToPlace, which are easy to read and would not occlude the consumer's sight while completing the tasks.

C CODEBOOK AND THEMES FROM QUALITATIVE DATA ANALYSIS

We used thematic analysis [25], and deductive and inductive coding [39] to analyze qualitative data, collected from preliminary needs-finding study (Sec. 3) and final user study (Sec. 8). As part of supplementary material, we attached the resultant codebook in Fig. 21, Fig. 22 and Fig. 23, respectively. Notably, "Count" refers to the number of quote for each theme or code. It is also possible that multiple codes are assigned to one quote.

D ETHICAL DISCLAIMER

This work has been approved by the Institutional Review Board (IRB). All Personal Identifiable Information (PII), such as the face

has been intentionally removed (*e.g.,* being pixelized or blurred) in this manuscript as well as all accompanion videos. Before each user study, we have obtained participants' consent on video and audio recordings, as well as heterogeneous behavior data collections. All participants have consented and acknowledged the data and results presented in this manuscript, to be published and presented publicly for research purposes. While monetary incentives were not provided, all participants had the opportunity to try out and experience state-of-the-art MR technologies, and to know further about our research.
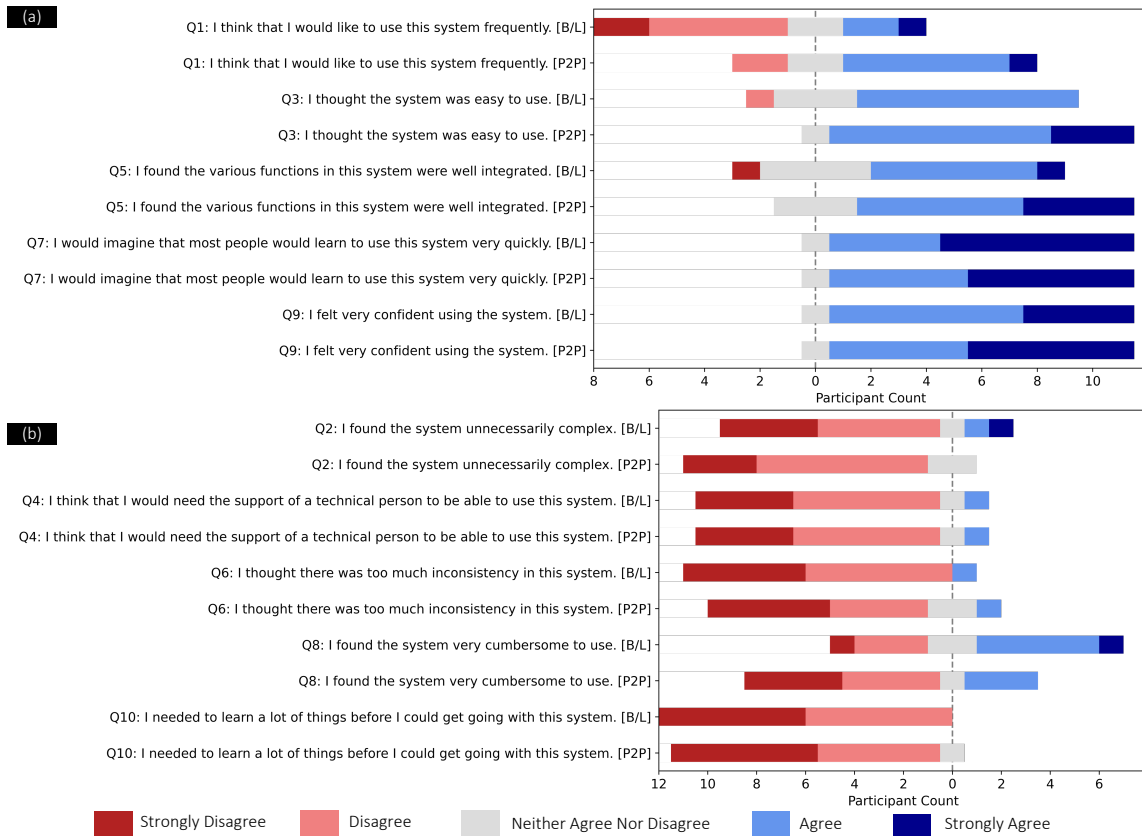
Figure 19: Survey results of SUS questionnaires of consumption pipeline evaluations. We use "B/L" to refer to the baseline interface, and "P2P" to indicate PaperToPlace, which delivers spatialized and context-aware instruction step. To increase readability, we cluster the survey results of positive statements (Q1, Q3, Q5, Q7, Q9) into subplot (a), where a higher level of agreement indicates a better user experience. The survey results of negative statements (Q2, Q4, Q6, Q8, Q10) are clustered into subplot (b), where a lower level of agreement indicates a better user experience.
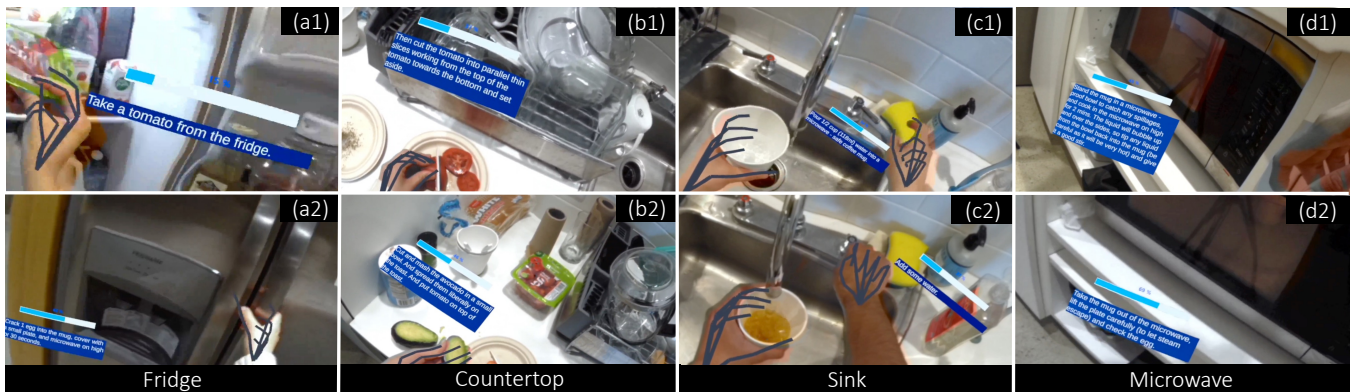
Figure 20: First-person view through MR of the examples of placing instruction step next to the key objects with our consumption pipeline. Example key objects include fridge (a), countertop (b), sink (c), and microwave (d).

Theme/Code	Count	Theme/Code	Count
Reading Instruction Document	64	**Thoughts while Designing MR based Instruction Experience**	44
Reading method	13	Needs of spatial guidance in MR	7
Overwhelm of irrelevant information	11	Unnecessary spatial guidance can be visual disturbance in MR experience	7
Challenge of finding relevant information	6	Segment instruction document and only keep relevant information	6
Frustration on document reading	4	Capable of revising existing instruction document while designing MR experience	4
Lack of relevant visual components in documents	3	Capable of hiding irrelevant information in MR experience	6
Lack of emphasize on key information	10	Needs to connect document content to the workflow of activities	6
Needs of personalized information	9	Virtual instruction should be closed to the user	4
Lack of understanding of key concept	2	Comments related to the visualizations of progress status	4
Lack of key information in the instruction	6		
Connecting to Environment and Real-World Activities	26		
Relying on common sense	9		
Needs some forms of feedback	1		
Referring to document repeatedly while doing tasks	2		
Failures of finding key objects or ingredient	2		
Hard to connect with real-world activities	12		

Figure 21: The codebook that resulted from our qualitative analysis of interview data for preliminary needs-finding study. "Count" refers to the number of quote for each theme or code. It is possible that multiple codes are assigned to one quote.

Theme/Code	Count	Theme/Code	Count
Overall Comments of the Pipeline	78	**Can ML Support an Easier Authoring Workflow?**	143
Feedbacks related to the system integrations	8	Comments related to overall performance	21
Benefits of using mobile camera on iPad to capture existing paper-based instruction document	3	Comments related to overall efforts	35
Simple and easy to learn	19	Helpfulness to the mental cognitive load and thought process	15
Benefits of segmenting the parsed instruction document	12	Accuracy of predicted key objects	32
Other miscellaneous setbacks and improvements	4	Experience and thought process while checking and correcting the predicted key objects	14
Other generic overall comments	32	Experience of real-time prediction of key objects upon the change of text of instruction steps	10
Can ML Support a Faster Authoring Workflow?	26	Affordances of visualization of confidence using color scale	14
Subjective experience of the temporal benefits	17	Other miscellaneous setbacks and improvements	2
Rationales of potential reasons for a faster authoring workflow with ML-supported mode	9	**Comments after Evaluating the Authored Instruction MR Experience**	31
		Positive comments and appreciations	25
		Setbacks and improvements	6

Figure 22: The codebook that resulted from our qualitative analysis of interview data for authoring pipeline evaluations. "Count" refers to the number of quote for each theme or code. It is possible that multiple codes are assigned to one quote.

Theme/Code	Count	Theme/Code	Count
Context Awareness Reducing the Effort for Context Switching	90	**Other Aspects of System Usability**	97
Subjective feelings of reduced efforts for context switching	17	System is easy to learn	7
Benefits and subjective experience of accuracy of placed instruction step	19	Overall affordance in terms of mental workload	15
Costs and setbacks of occasional non-optimal placed instruction step	16	Overall affordance in terms of task completion time	7
Methods to requesting instruction step through hand pinch	22	Benefits for helping with spatial understanding	14
Promising and have potentials for more complex and spatialized tasks	15	Needs of other types of metadata	2
Imperfect for steps that are involved with find-grained activities	1	Occlusions of rendered skeleton of tracked hands	5
Segmented Instruction Helps Findings the Relevant Information Easier	48	Quality of video see-through (such as visual distortion *etc.*)	14
Subjective feelings of more calmness	6	Failure tracking of hand and controllers	1
Other comments of merits of segmenting instructions	24	Needs of hand(s) for baseline interface	18
Comments related to readability and monolithic nature of the baseline interface	18	Other miscellaneous comments	14
		Brainstorm and Discussions related to Broader Applications	31
		Enumerate examples of possible applications	18
		Needs of multimedia support	1
		Benefits of deciding optimal placements of instruction step for the discussed applications	8
		Improvements of current system for future applications	5

Figure 23: The codebook that resulted from our qualitative analysis of interview data for consumption pipeline evaluations. "Count" refers to the number of quote for each theme or code. It is possible that multiple codes are assigned to one quote.