International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

MultiUX: A Human-AI Collaborative Tool to Facilitate Multiple Usability Test Video Analyses

Luyao Shen, Qing Shi , Emily Kuang, Linjie Qiu , Shixu Zhou , Pan Hui & Mingming Fan

To cite this article: Luyao Shen, Qing Shi , Emily Kuang, Linjie Qiu , Shixu Zhou , Pan Hui & Mingming Fan (02 Jan 2026): *MultiUX*: A Human-AI Collaborative Tool to Facilitate Multiple Usability Test Video Analyses, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2025.2605526

To link to this article: https://doi.org/10.1080/10447318.2025.2605526

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2025.2605526

MultiUX: A Human-AI Collaborative Tool to Facilitate Multiple Usability Test Video Analyses

Luyao Shen[a], Qing Shi[a], Emily Kuang[b], Linjie Qiu[a], Shixu Zhou[c], Pan Hui[a,c] and Mingming Fan[a,c]

[a]Computational Media and Arts, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China; [b]Electrical Engineering and Computer Science, York University, Toronto, Canada; [c]Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong

ABSTRACT

Analyzing multiple usability videos is essential for identifying common problems and prioritizing them. While existing tools assist in identifying problems within single videos, the manual process of comparing and categorizing them remains time-consuming. Inspired by prior work that suggested a human-AI collaborative approach could improve the efficiency and completeness of analyzing single videos compared to either humans or AI alone, we extended this approach to investigate its effectiveness in analyzing multiple videos. We designed *MultiUX* to assist in strategically analyzing multiple videos by comparing and categorizing problems across them, and optimizing the analysis sequence through video recommendations. *MultiUX* was compared to a baseline in a between-subjects study involving 20 UX evaluators. Results showed that *MultiUX* supported analyzing more videos and identifying more common problems that were encountered by more users in a given time. Additionally, *MultiUX* guided participants to analyze multiple videos more strategically.

KEYWORDS

User experience; usability analysis; multiple video analysis; artificial intelligence; visual analytics

1. Introduction

Usability testing is a widely used method for identifying usability issues users encounter when interacting with a product or system (Fan et al., 2020b). During these tests, users follow think-aloud protocols, verbalizing their actions and thoughts as they perform specific tasks using the product (Nørgaard & Hornbæk, 2006). Research has shown that testing with around ten users can uncover approximately 80% of usability problems (Hwang & Salvendy, 2010). Thus, user experience (UX) evaluators often collect multiple usability videos in a single session to identify common problems faced by different users. Analyzing these videos provides valuable insights into the severity of issues, helping prioritize redesign strategies effectively (Georgsson & Staggers, 2016; Jeddi et al., 2020; Nielsen, 2022). The analysis of usability problems in multiple videos typically involves identifying issues in each video, then comparing and categorizing these problems across all videos to find commonalities (Kjeldskov et al., 2004). Standardized frameworks, such as usability problem taxonomies (Keenan et al., 1999) and usability heuristics (Nielsen, 2024), guide this categorization process. However, these frameworks may not fully address the diverse range of problems encountered in complex or state-of-the-art products (Khajouei et al., 2011; Liljegren, 2006). As a result, evaluators often employ a bottom-up clustering method, where multiple evaluators independently derive and discuss categories until they reach agreement (Delice & Güngör, 2009; Petrie & Power, 2012). While this approach provides more tailored insights, it is time-consuming and requires extensive collaboration. Additionally, time constraints during analysis may lead to missed issues or misinterpretations (Kuang et al., 2022, 2023; Soure et al., 2022). Therefore, analyzing multiple videos is still a resource-intensive, challenging process (Følstad et al., 2012b; Kjeldskov et al., 2004).

CONTACT Mingming Fan mingmingfan@ust.hk Computational Media and Arts, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

To address these challenges, previous studies have applied machine learning techniques to extract video features such as loudness and pitch, indicating potential usability problems (Batch et al., 2024; Fan et al., 2020c; 2022; Soure et al., 2022). Moreover, Kuang et al. (2024) utilized transcriptions of think-aloud content and instructed large language models (LLMs)[1] to identify usability issues based on the transcripts. This allows UX evaluators to efficiently annotate problems in individual videos. However, the process of comparing and categorizing problems across multiple videos remains underexplored.

Our study aims to explore how technology can support the comparison and categorization of problems, thus supporting multiple usability video analyses. While LLMs have been successfully applied to analyze qualitative data, such as semi-structured interview transcripts and usability protocols (Burnam, 2023; De Paoli, 2024; Gao et al., 2024; Kocaballi, 2023; Singh et al., 2024), their potential to assist in video analysis remains underexplored. Inspired by these advancements, we propose integrating LLMs to identify and categorize usability problems by analyzing the transcripts of multiple videos.

To effectively leverage LLMs in analyzing multiple usability videos, it is crucial to understand the challenges and specific needs of UX evaluators. Our formative study showed that UX evaluators face challenges in managing and categorizing numerous usability problems, assessing their frequency across multiple videos, and optimizing video analysis under time constraints. Additionally, they struggle with managing data across different platforms. These challenges highlight the need for an integrated platform that facilitates problem categorization, frequency assessment, and analysis sequence optimization in multi-video usability testing.

Building on these findings, we introduced key features to facilitate multiple usability video analyses: 1) AI-assisted problem identification and categorization to recommend similar problems within the same category, 2) a set of categories inferred by AI and UX evaluators to manage and filter problems, 3) video recommendation based on problem similarity to optimize the analysis sequence, and 4) summary visualization to display problem distributions across all videos based on categories, facilitating data revisitation. Incorporating these features, we designed *MultiUX* to support multiple usability video analyses.

To understand how *MultiUX* and its key features support the determination of the analysis sequence, problem comparison, and categorization in multiple videos, we developed a baseline version that only helped identify the problem within single videos, removing these features. This baseline design allowed us to understand the effectiveness of the key features and how the evaluators used them in multiple video analyses. The baseline mirrors common tools used by UX evaluators, which typically provide potential usability problems in natural language without the ability to compare problems across multiple videos. We conducted a between-subjects study involving 20 UX evaluators. Each participant analyzed ten usability videos of one product in a limited time. Our results showed that *MultiUX* enhanced multiple video analyses by improving efficiency and refining the analytical process compared to the baseline. Specifically, using *MultiUX*, UX evaluators were able to review more videos and identify a higher number of common problems encountered by multiple users within the same timeframe. Furthermore, *MultiUX* strategically guided participants in determining the sequence of video analysis based on recommendations. In summary, our contribution encompasses:

- A formative study that uncovered the practices and associated challenges UX evaluators faced when analyzing multiple usability videos.
- *MultiUX*, a system that supports multiple usability video analyses by identifying and categorizing usability problems, recommending the sequence of videos, and summarizing problems in an interactive visualization.
- A user study that examined how the key features of *MultiUX* assisted participants in identifying a broader and more comprehensive set of usability problems encountered by a wider range of users.

2. Related work

2.1. Identifying and categorizing usability problems across multiple videos

Usability testing is the most common method to evaluate usability when developing various types of products (Fan et al., 2020b; McDonald et al., 2012). Think-aloud protocols, often regarded as the primary method of usability testing, are widely used by UX professionals to obtain insights. In these

sessions, users verbalize their thought processes as they complete the tasks involved in the usability test. (Nørgaard & Hornbæk, 2006). Prior research found that detecting 80% usability problems required about ten test users (Hwang & Salvendy, 2010). Thus, UX evaluators collect multiple usability videos in the testing session for further analysis, including problem identification and categorization.

The frequency of problems in multiple videos is a critical factor in usability analysis (Sauro, 2014), useful for determining the severity (Georgsson & Staggers, 2016; Jeddi et al., 2020; Kuang et al., 2022; Nielsen, 2022). Researchers have developed several standardized categories to guide the categorization of usability problems and count the frequency, such as the usability problem taxonomy (Keenan et al., 1999), usability heuristics (Nielsen, 2024), and ISO definition (Liljegren, 2006). Although these categories have been leveraged in specific fields, such as medical-related products (Georgsson & Staggers, 2016; Jeddi et al., 2020; Liljegren, 2006), previous research has highlighted that the increasing complexity and diversity of products, along with the varied user behavior across these products, make these categories insufficient due to their incompleteness, non-exclusivity, and lack of specificity (Khajouei et al., 2011; Liljegren, 2006).

Besides existing categories, another method is the bottom-up clustering of usability problems to summarize self-defined categories (Delice & Güngör, 2009; Petrie & Power, 2012). For example, identifying general usability problem categories and subcategories through various websites (Petrie & Power, 2012) and categorizing based on product functions (Marcio Silva et al., 2014). However, this process is usually time-consuming and labor-intensive, necessitating at least two UX evaluators to first independently review multiple videos, identify and categorize problems, and then achieve consensus (Borlinghaus & Huber, 2021; Følstad et al., 2012a; 2012b; Khajouei et al., 2011; Kuang et al., 2022). These challenges emphasize the need for a method that not only supports problem identification but also enhances efficiency in comparing and categorizing problems. Additionally, it should be adaptable to the evolving complexity of the problem categories.

2.2. AI-based multimodal video analysis tools

We live in a multimodal world where video has become the predominant form of media. However, manually processing the vast amount of video content remains labor-intensive and time-consuming. Given the remarkable capabilities of LLMs in language and multimodal tasks, they have recently been applied to assist in recognizing and interpreting video content (Tang et al., 2025). Visual information serves as the primary modality for understanding videos, such as answering user queires by extracting and analyzing keyframes (Wang et al., 2025). To further enhance video understanding, researchers have begun integrating additional modalities. For example, aligning textual narrations with visual information improves comprehension of temporally sensitive content such as instructional videos (Chen et al., 2025). Similarly, combining subtitles and audios with visual information has been shown to improve the accuray of video-related question answering (Fu et al., 2025). However, these advances in multimodal video understanding primarily focus on algorithmic optimization and are typically limited to the analysis of individual videos, neglecting cross-video comparisons and synthesis.

Analyzing multiple videos is a common activity in many scenarios, such as evaluating individual performance in group activities using multiview videos or understanding a surgical procedure by watching videos performed by different surgeons (Charoenkulvanich et al., 2019; Xia et al., 2014). The analysis process typically involves segmenting videos based on spatial, temporal, or user events, and then categorizing these segments for focused comparison and analysis. For example, Charoenkulvanich et al. analyzed individual and joint actions by detecting hands in multiple videos captured from each person in group tasks (Charoenkulvanich et al., 2019). Similarly, Xia et al. designed a multi-video interaction to support simultaneous viewing of multiple videos of the same surgical procedure performed by different surgeons, allowing users to identify common features (Xia et al., 2014). Prior studies have developed automatic tools to segment multiple videos, primarily by detecting specific visual elements such as hands (Charoenkulvanich et al., 2019) or buildings (Wang et al., 2014). However, although these tools grouped video segments that contain the same visual element, users still needed to manually annotate them and define categories, which is time consuming and labor intensive (Hagedorn et al., 2008; Wang et al., 2018).

Video segmentation approaches based on specific visual element detection demonstrate limited applicability in usability video analysis, as they primarily focus on identifying UI elements. However, such an approach fails to effectively distinguish between multiple usability problems that may arise during a single UI interaction session, thereby constraining the granularity of segmentation in usability evaluation scenarios. While recent progress in multimodal and single-video understanding has been substantial, research on tools that enable analysis across multiple videos remains relatively sparse and less mature. This gap motivates the exploration of new approaches to facilitate more efficient and strategic multiple video analysis.

2.3. AI-based qualitative coding

Qualitative coding plays a crucial role in usability analysis, as it helps UX evaluators organize and interpret raw testing data (e.g., think-aloud transcripts) into structured findings. To improve the efficiency of qualitative coding, prior studies have utilized machine learning to partially automate this process (Gebreegziabher et al., 2023; Lu et al., 2024; Rietz & Maedche, 2021). These tools can apply human-defined codes to all data (Liew et al., 2014b; Marathe & Toyama, 2018) or summarize codes by learning from a small set of manual annotations (Gebreegziabher et al., 2023; Liew et al., 2014a; Rietz & Maedche, 2020, 2021). Moreover, researchers found that these tools struggled to summarize abstract codes like "atmosphere" and were further limited by categorizing based solely on similar words without semantic understanding (Borlinghaus & Huber, 2021; Rietz & Maedche, 2021, 2020). With the capabilities of capturing semantic and latent meanings of text, LLMs showed data coding quality similar to that of humans (Dai et al., 2023; De Paoli, 2023; Schiavone et al., 2023; Xiao et al., 2023). LLMs can automatically generate codes and themes for qualitative data (Qiao et al., 2025; Singh et al., 2024). Nevertheless, given the inherently reflexive and interpretive nature of qualitative coding, LLMs often lack the hermeneutic depth required for developing nuanced codebooks. To address this limitation, recent research has proposed human-AI collaborative method in which human coders design and refine the codebook, and LLMs subsequently adapt and apply it to large-scale datasets (Dunivin, 2025). Beyond text-based data, recent work has extended LLMs to video content analysis by converting videos into multiple textual representations, such as video descriptions, summaries, and on-screen text. These textual representations are then used as input for LLMs to generate annotations, which are subsequently validated and refined by human experts (Ghosh et al., 2025). However, these studies primarily focus on assessing LLM performance across datasets, rather than designing or evaluating tools that support human-AI collaborative coding processes.

Several studies have recently developed human-AI collaborative tools for different stages of qualitative coding. For example, SenseMate enables users to review and react to AI-recommended codes for selected text segments, while ThemeViz allows users to explore multiple AI-generated theme versions that align with the iterative nature of theme development process (Kang et al., 2025; Overney et al., 2024). Similarly, CollabCoder leverages LLMs to support human-human collaborative qualitative coding by suggesting ways to group semantically similar codes proposed by different coders (Gao et al., 2024). Together, these studies demonstrate the potential of human-AI collaboration in comparing and categorizing data from multiple sources, suggesting their applicability to categorize usability problems across multiple videos. However, although bare LLMs (i.e., default or fine-tuned models used directly for video content analysis without integration into interactive tools or practical workflows) have been explored, existing human-AI collaborative coding tools remain limited to text-based data and overlook the unique challenges of multimodal video analysis.

2.4. Human-AI collaborative tools for usability video analysis

Analyzing usability videos is challenging, as UX evaluators must listen to what users say while watching what they do simultaneously to identify usability problems. Moreover, UX evaluators are required to analyze these videos in a limited time in practice, potentially leading to the oversight of crucial information (Fan et al., 2020c; Følstad et al., 2012a; Kuang et al., 2022). Facing this challenge, AI is introduced as an assistant to provide another perspective to the human UX evaluator.

Prior studies have leveraged machine learning to identify verbalization features (e.g., low speech rate, silence, and question words (Fan et al., 2019; Soure et al., 2022)) that are correlated with usability problems (Fan et al., 2019, 2021, 2020a; Fan & Zhu, 2021). These verbalization features are pre-processed and integrated into web-based analytical tools to assist in highlighting video segments with potential problems and improving the efficiency of the analysis (Batch et al., 2024; Fan et al., 2020c; Soure et al., 2022). ChatGPT has been utilized in the form of conversational AI assistants to analyze usability videos alongside human UX evaluators. It suggests potential usability problems in natural language and has proven effective in identifying more comprehensive problems more efficiently within single videos (Kuang et al., 2024). However, prior efforts have mainly examined AI assistance in the context of individual videos, focusing on feature detection or single-video interpretation. In contrast, our work expands the capabilities of these tools to support the comparison and categorization of problems across multiple videos, and further investigates how such expansion influences human evaluators' multiple video analysis strategies and practices.

3. Design of MultiUX

We conducted semi-structured interviews with twelve UX evaluators (11 females and 1 male) with two primary goals: first, to gain insight into the current practices and challenges in analyzing and comparing multiple usability videos; and second, to identify their requirements for a tool that can assist them in multiple usability video analyses. Each interview lasted about 30 min. During the interview, participants were asked to respond to questions based on their daily work experiences. All participating experiments of this work passed the ethical review of the Hong Kong University of Science and Technology (Guangzhou) with review approval number HSP-2025-0003.

We recruited participants by distributing posters on social media platforms. Eligibility criteria included having experience with at least two usability analysis projects. We used a pre-study survey to gather participants' demographic information. We also collected their familiarity with usability analysis using a 5-point Likert scale, where 1 corresponded to "extremely unfamiliar" and 5 to "extremely familiar.". Overall, these participants work in the software, gaming, and car industries. Their average age is 28.9 ($SD = 3.6$). On average, they had three years of experience in UX ($SD = 1.9$). Eight (66.7%) of them reported being very familiar or extremely familiar with usability analysis, while the remaining four indicated moderately familiar ($M = 4, SD = 0.87$).

The interviews were conducted remotely using an online meeting tool. The tool automatically transcribed the interview data, which was then reviewed and corrected by a researcher. Two researchers analyzed the transcripts using thematic analysis (Rosala, 2024), following the process of generating codes, grouping them into common themes, and discussing resolving conflicts (Braun & Clarke, 2006).

3.1. Design considerations

Based on insights from these interviews and prior work, we derived four design considerations (DCs) to guide the design of *MultiUX*, a tool that supports UX evaluators to analyze multiple usability videos.

DC1:Facilitate managing and categorizing problems across multiple videos. When reviewing usability videos, all UX evaluators (N = 12) recorded timestamps and corresponding problem descriptions in a text editor. Subsequently, they engaged in a problem categorization process to assess the frequency of each category. This process brought several challenges. P2 and P3 noted that "it was hard to find similar ones in a long list of problems." P9 shared her experience using a commercial application Lookback (Lookback, 2024) to analyze multiple videos, noting that problems were associated with the corresponding videos, causing difficulty recalling problems from earlier videos after switching multiple times. P3 noted that while labeling each problem could aid in categorization, it introduced new complexities in category management and updating. Standardized categories, like Nielsen's heuristics, are too high-level to use in the work context. Several participants (N = 9) expressed a preference for using self-defined categories, which often lack sufficient descriptiveness and require frequent updates throughout the analysis process.

DC2: Enhance efficiency in counting and updating the frequencies of usability problems. To reduce the time needed to evaluate the problem severity, several participants (N = 3) assessed problem frequencies according to their memory, a method susceptible to memory bias. An additional challenge emerged during discussion scenarios that involved deciding the category of usability problems. P10 described, "Frequent recategorization leads to a constant recounting of frequencies, increasing our effort." Faced with these challenges, participants expressed a need for features that could streamline the process of assessing frequencies, such as automatically counting and building a table to filter problems according to frequencies.

DC3: Assist UX evaluators in determining the analysis sequence while maintaining track of the viewed video segments. UX evaluators often conduct usability tests with about ten users to identify a wider range of usability problems (Hwang & Salvendy, 2010; Mugunthan, 2023). Given time constraints, they often cannot analyze all the videos thoroughly. P12 shared her strategy for selecting representative videos, "I would quickly go through each video and identify representative users who encountered the widest range of problems. Then, I would prioritize the detailed analysis of these videos. However, this approach to selecting videos was time-consuming." After an in-depth analysis of one or two randomly selected videos, some UX evaluators shifted their strategy to identify common usability problems encountered by multiple users, requiring frequent video switches. P10 frequently switched between videos to analyze a specific website, reviewing segments from multiple videos that cover the target page to count users who encounter problems. These frequent switches can result in disorientation, losing track of the viewed segments of the videos.

DC4: Support the display and usage of different types of data generated during the analysis. A prevalent practice among UX evaluators (N = 8) involved using multiple applications to view videos and take notes, in line with previous survey findings (Følstad et al., 2012a). UX evaluators devoted extra time to cutting video segments and storing them along identified problems to reduce revisiting video players and facilitate sharing findings. P3 and P12 highlighted the effort required to screenshot the interface with problems. In addition to linking video segments or screenshots to their original videos, the evaluators stressed the display of additional data such as problem descriptions, categories, and user demographics (P3,5). The display of this information was beneficial in three scenarios: reviewing detailed actions and verbalizations of users who encounter common usability problems (P6), discussing the severity of usability problems with stakeholders (P3,9,12), and rechecking the product usability report (P9).

4. MultiUX system

We developed *MultiUX* based on the aforementioned DCs. Our main goal was to improve the efficiency and confidence of UX evaluators in analyzing multiple usability videos. In this section, we first describe *MultiUX*'s key features that facilitate multiple video analyses. Following this, we present the implementation of *MultiUX* and the quality assessment of AI-generated suggestions.

4.1. Key features of the MultiUX

4.1.1. MultiUX interaction paradigm

As shown in Figure 1, *MultiUX* was developed as a visual analytics interface with timeline visualization. The selection of the interaction paradigm was informed by a review of interface designs within the same application domain. We identified two primary interface paradigms: visual analytics tools and conversational AI assistants. The detailed advantages and disadvantages of these paradigms are explained in Appendix A.

Given that the primary design goal of *MultiUX* is to identify and categorize usability problems in multiple videos using a category corpus and optimize the analysis sequence, a visual analytics interface is more suitable. It offers better organization by distributing information across multiple panels, including a timeline visualization supporting similar problem recommendations (**DC2**), a category panel for managing and categorizing problems (**DC1**), and a video recommendation panel to guide the analysis

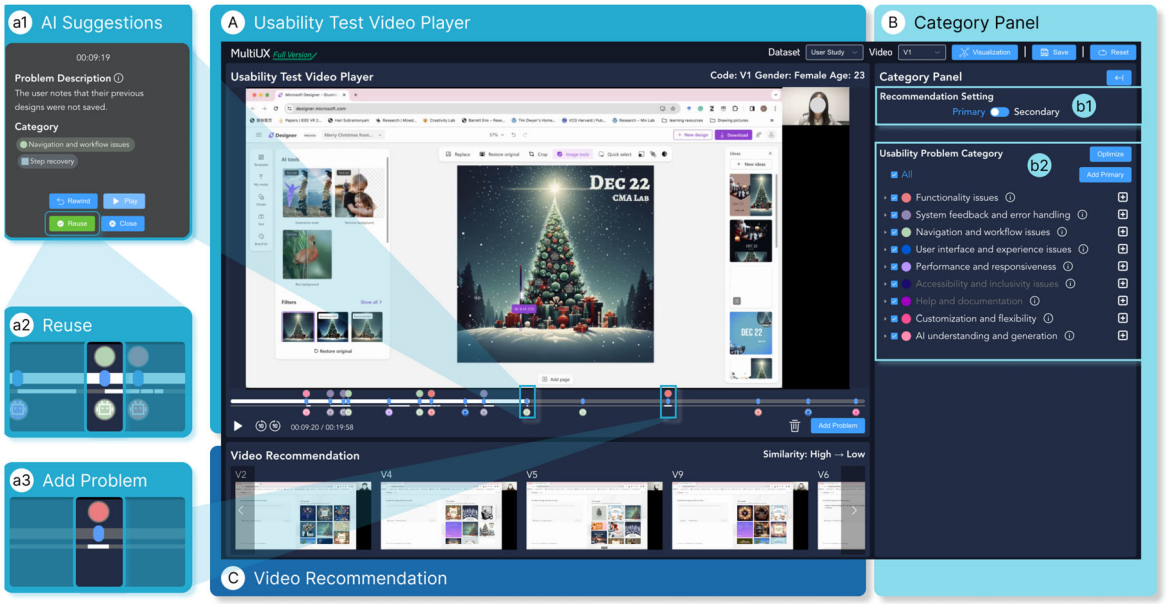

Figure 1. *MultiUX* Interface, showing the scenario of analyzing ten videos of an AI-powered design product: (A) *usability test video player* for viewing the video and identifying usability problems; (B) *Category panel* for displaying both the AI-inferred and manually added categories; (C) *Video recommendation* is based on the similarity.

sequence (**DC3**). Furthermore, *MultiUX* records and displays all data during the analysis process (**DC4**). This interface also allows for scalability, accommodating future interactive features.

4.1.2. Problem categorization and recommendation in the video timeline

As shown in Figure 1A, a video player is incorporated to streamline the video review process (**DC4**). This player supports standard functionalities such as play, pause, forward, and rewind. The icons for manually added problems and AI suggestions are placed above and below the timeline, respectively.

AI suggestions pop up at identified timestamps (Figure 1,a1) during the first playback, helping UX evaluators quickly record potential usability problems. UX evaluators can "Reuse" AI suggestions to add problems or "Close" them if irrelevant. After clicking "Reuse", an icon appears above the timeline at the same timestamp (Figure 1,a2), indicating agreement by UX evaluators. Additionally, *MultiUX* allows modifications to AI suggestions to match UX evaluators' work habits (Figure 2A,B). The section titled *Same Usability Problem* (Figure 2C) catalogs timestamps of problems within the same category in all videos. UX evaluators can easily switch between videos to analyze problems within the same category (**DC1**) and assess the frequency of these problems (**DC2**). Blue links denote completed analyses. When identifying the problems overlooked by AI, UX evaluators can "Add Problem" to note them down. After adding the problem description, UX evaluators can choose categories from the existing set or get recommendations from ChatGPT. An icon is then displayed above the timeline at the corresponding timestamp (Figure 1,a3).

4.1.3. Human-AI Co-created usability problem category set

To explore the preferences of UX evaluators for the level of detail in the categorization of usability problems, we integrated a two-tier categorization system in *MultiUX*. As shown in Figure 1B, *MultiUX* initially displays AI-inferred categories derived from all AI-identified problems. We chose circles to represent categories and squares to represent subcategories, with each color corresponding to a category. The categories and subcategories can be added by clicking "Add Primary" and respectively. Hovering on allows users to learn about the definition of the category. The "Optimize" function can analyze existing categories via ChatGPT and provide suggestions to merge overlapped ones (Figure 3B), making the set of categories more distinct and systematic. Both newly created or merged categories in the usability problem category and time-stamped annotations are updated simultaneously (**DC1**). In addition, UX evaluators can select one or more categories of interest from the panel (**DC2**), leading to

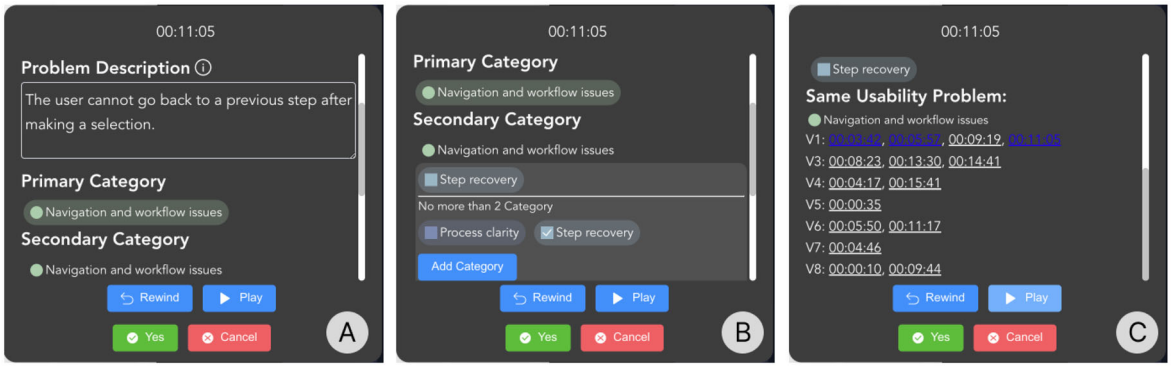

Figure 2. Pop-ups displaying AI suggestions: (A) refining problem description; (B) Modifying problem categories; (C) Displaying usability problems within the same category.

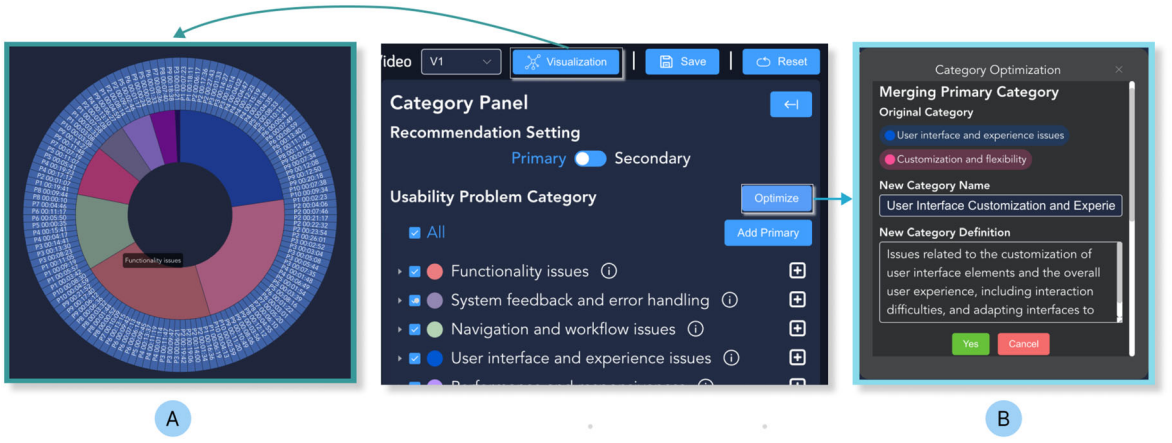

Figure 3. User interfaces illustrating: (A) summary visualization; (B) A pop-up for optimizing categories.

a recalculation of similarity and adjustment of the order of the video recommendation and changing the display information in the visualization. By default, *MultiUX* only displays categories and UX evaluators can click in the *Recommendation Setting* to expand subcategories. Information about subcategories will also be displayed in the video timeline and *Summary Visualization*.

4.1.4. Video recommendation based on identified problem similarities

MultiUX employs the Jaccard similarity coefficient as a metric to calculate the similarity between usability videos (Sun et al., 2011). The usability videos in the recommendation list (Figure 1C) are organized in descending order according to their similarity coefficient (**DC3**). The calculation formula is as follows:

$$Sim(u,v)^{Jaccard} = \frac{|I_u \cap I_v|}{|I_u \cup I_v|}$$

where I_u and I_v are the sets of usability problem categories of Video u and Video v respectively. Furthermore, *MultiUX* allows UX evaluators to select one or more specific categories from the *Usability Problem Category*, enabling them to conduct focused analysis. Incorporating video and usability problem recommendations within *MultiUX* can allow UX evaluators to navigate between various videos and distinct timestamps within the same video. We implemented a narrower timeline below the main one (Figure 1A) to enhance user orientation during this navigation process (**DC3**).

4.1.5. Summary visualization of identified problems by category across multiple videos

As shown in Figure 3A, we selected the sunburst diagram as the summary visualization in our system. Each arc inside this sunburst diagram corresponds to a category, with each outer arc corresponding to a problem, displaying the original video and timestamp of the problems. The overview mode of the

sunburst view diagram displays all problems across multiple videos, highlighting the ones processed by UX evaluators, while the filter mode shows the summary of the selected video. Clicking on a category arc allows zooming in, while clicking on the central circle enables zooming out (**DC2**). Additionally, clicking on the timestamp arc can preview and jump back to the corresponding video (**DC4**).

We selected sunburst diagrams for their effectiveness in visualizing hierarchical information and efficiency in conveying proportion. Additionally, this radial visualization offers scalability through interactive design elements, such as transitions between focused and overview regions (Stasko & Zhang, 2000).

4.2. Implementation of MultiUX offline pre-processing, real-time interaction, and interface design

Users are required to upload the set of usability test videos along with their corresponding transcripts to *MultiUX*. In addition, product UIs need to be provided (Figure 4). We used the ChatGPT − 4 model to implement both offline pre-processing and real-time interaction (OpenAI, 2024a).

4.2.1. Offline pre-processing implementation

Upon receiving user-uploaded data, the subsequent processing is automated. Given that ChatGPT − 4 accepts text or image inputs, and previous studies have confirmed its effectiveness in identifying usability problems using test transcripts (Kuang et al., 2024; OpenAI, 2024a), we used transcripts and screenshots of the product's primary interfaces to enhance ChatGPT's contextual understanding. We adopted a strategy based on principles of prompt engineering, specifically the guideline to split complex tasks into simpler subtasks (OpenAI, 2024b). Accordingly, we instructed ChatGPT to process the transcripts through the three steps shown in Figure 4 *AI offline pre-processing*. One challenge encountered was the variability in ChatGPT's output. To address this, we instructed ChatGPT to identify problems through three separate chat threads for each video and then synthesize these results (Ronanki et al., 2022). This approach was similarly applied when inferring problem categories. For the step of *assigning category labels to each problem*, we adopted the strategy of selecting the category most frequently assigned from three rounds for each problem. We also instructed ChatGPT to provide justifications for its problem categorizations in each round.

The prompt template used in this process followed the structure [Background Information] + [Task Description]. An example of [Background Information] is provided below:

> The following is the transcript of a user study where a participant used the think-aloud protocol to complete the following tasks on an AI-powered design website: Design a Christmas party poster with date, time, location, and activity information. The transcript contains timestamps and the words spoken by the participant. To increase your contextual understanding, I uploaded screenshots of the main interfaces of this website."

The [Task Description] for each step is presented in Table 1.

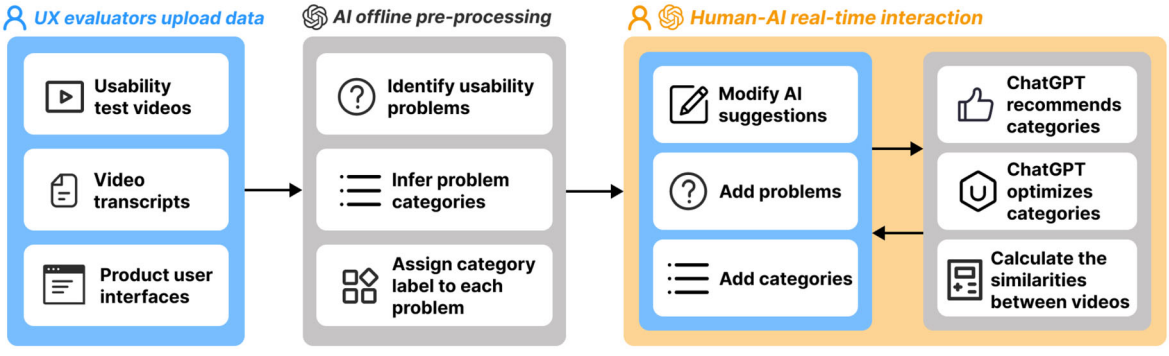

Figure 4. *MultiUX* Is a human-AI collaborative tool designed to assist UX evaluators in analyzing multiple usability test videos. Uploaded data undergoes a three-step AI preprocessing phase before being provided to UX evaluators. UX evaluators can then modify and complement the AI-generated data. Based on evaluators' inputs, the AI provides real-time support.

Table 1. Given to the ChatGPT for each step.

Offline pre-processing step	Task description
Identify usability problems	Based on the transcript and screenshots, can you identify which usability problems the participant may have encountered and when these problems occurred? Provide your response in the format: Timestamp, Problem description.
Infer problem categories	Can you give me a list of categories and subcategories of all identified usability problems? Each category can be divided into several subcategories. Please add a definition for each category. Provide your response in the format: Category, Definition, Subcategory, Definition.
Assign a category label to each problem	Can you analyze each usability problem and assign a category and a subcategory you find most appropriate? Provide explanations for your assignment. If you think that the categories in the provided list do not match, you can add new ones. Remember to add definitions when adding new categories. Provide your response in the format: Category, Subcategory, Explanation.

Table 2. Task Description Provided to ChatGPT for each real-time task.

Real-time task	Task description
Recommend a category for an input problem description	First, review the JSON file named [problem_category_set] to ensure you understand the names and definitions of all categories and subcategories. Based on the [user input] of the usability problem description, recommend the most suitable category and subcategory for this problem. Explain your recommendation. Provide your response in the format: Category, Subcategory, Explanation.
Optimize usability problem categories	Review the JSON file named [problem_category_set] to ensure you understand the names and definitions of all categories and subcategories. First, analyze all subcategories to determine if any similar subcategories can be merged, if any missing subcategories need to be added, or if any unclear subcategory names and definitions require refinement. After optimizing the subcategories, analyze all categories following the same steps. Finally, output all optimization suggestions in the following format: Recommended action, Optimized category name, Optimized category definition.

After processing, we compile a set of data for each video, including a list of identified problems, along with associated timestamps, categories, detailed descriptions, and ChatGPT explanations. These AI suggestions were displayed in *MultiUX* to provide a preliminary analysis for UX evaluators.

4.2.2. Real-time interaction implementation

As illustrated in Figure 4, the AI facilitates real-time interaction in response to UX evaluators' actions, including modifying and complementing AI suggestions. AI can recommend categories for manually added problems and optimize problem categories developed through human-AI collaboration. Moreover, *MultiUX* can recommend videos by calculating similarities based on problem categorization. The task descriptions are detailed in Table 2.

4.2.3. Interface design implementation

We implemented *MultiUX* as a web application and deployed the tool and the videos on Tencent Cloud to enable access to the tool remotely and robustly through a URL. *MultiUX* includes a back-end Django server and a front-end UI based on Vue.js and the videojs player library. The Django can connect to the ChatGPT API to handle requests sent back from the front end. We place problems and their associated categories in a JSON file, while storing the set of categories with definitions in another JSON file. We dynamically visualized these data on the front-end user interface in real time. The activity history of the video player (e.g., play, pause, rewind) and the UX evaluators' actions (e.g., switch video, reuse AI suggestions, add problems) were saved into the JSON format log data.

4.3. Quality assessment of ChatGPT-generated problems and categories

Based on surveys of HCI-related practices in both academia and industry, UX evaluators usually recruited approximately 10 users for usability testing (Hwang & Salvendy, 2010; Nielsen, 2012). Therefore, we recruited 10 users (6 females, and 4 males) and collected usability videos in which they used digital products following the think-aloud protocol (Hwang & Salvendy, 2010). Users were instructed to design a Christmas party poster with the date, time, location, and activities on an AI-powered design website. These 10 users were labeled from U1 to U10.

To further evaluate the generalizability of *MultiUX* beyond the website product type, we additionally tested it on two other product types, a mobile app and a VR app, each with five usability test videos. The results, presented in Appendix B (Table B1), showed consistent trends in precision, recall, and categorization alignment, suggesting that *MultiUX* can effectively support usability analysis across different product types.

4.3.1. Quality assessment of ChatGPT-generated usability problems

We invited two UX evaluators (with an average of five years of UX experience) to first independently analyze ten videos and then collaborate to compile a consensus list of 127 usability problems, which served as the ground truth. The same ten videos were uploaded to *MultiUX* and 128 usability problems were identified. Upon comparing these problems with the ground truth, we found that although the precision was relatively low (20%) for one video, the overall precision (78. 1%) and the recall (78. 7%) were reasonable. The lower precision was shown in a video with only two identified usability problems where *MultiUX* detected ten, highlighting a potential tradeoff between precision and scalability in AI-driven analyses (Li et al., 2023; Van Berkel et al., 2022).

We examined the AI errors and found that most false positives were caused by the following three reasons. The most frequent cases occurred when the AI failed to distinguish design reasoning or editing behaviors from actual usability breakdowns. For example, U4 demonstrated his design thinking during the poster design process, noting that "I plan to add some color elements and keep the whole poster to two fonts. Thereby, I…". The AI over-inferred difficulty and misinterpreted this as "suggesting that the system might not be providing efficient tools or shortcuts for expert users to speed up the interaction." Another common reason was the complexity of simultaneously accounting for both the visual and audio channels of usability videos. A typical example is that after stating the design goal, U8 began typing the prompt silently and only said "OK. I click the Generate button" afterward. The AI overlooked the on-screen prompt input and incorrectly interpreted this silence as a sign of difficulty in finding the button. In other cases, users encountered a problem but resolved it quickly, yet AI failed to detect the recovery. For example, U5 said "But this layout isn't what I expected. There's no place to add activities," but immediately continued "I'll just add the activities elsewhere then." AI overlooked the resolution and still identified a problem, labeling it as "Inflexible design layout that does not accommodate all the required information".

Regarding the false negatives, the most frequent cases occurred when users expressed difficulties indirectly or with semantically softened phrases. For example, U4 remarked that "The fonts were a little ugly," and U5 said "The generated poster was not that playful," which AI did not interpret as usability problems. However, human evaluators considered these comments as indicating "limited font options" and "unsatisfactory generation results". Another source of false negatives was the AI's limited temporal sensitivity. When users complained about the same feature multiple times, AI mentioned it only once, whereas human evaluators recorded each occurrence. For example, U2 complained about the "Remove background" feature at timestamps 14:00 and 16:56, but AI only labeled the first instance. Finally, some errors were again due to the insufficient integration of visual and audio information. For example, U1 tried to mirror an image but failed, only saying "Mirror the image" during the attempt, which prevented AI from recognizing the problem from the audio data alone.

Overall, these AI errors highlight the complexities of identifying usability problems. Effective identification requires distinguishing users' design reasoning from actual usability breakdowns, recognizing whether actions represent trial-and-error or genuine difficulties, and intergrating multiple data sources, such as the verbal, behavioral, and contextual data to make accurate judgements. This complexity indicates that AI requires richer contextual inputs, domain-specific knowledge, and ongoing human oversight to adapt and correct errors that occured in such nuanced scenarios.

4.3.2. Quality assessment of ChatGPT-generated problem categories and ChatGPT's assignment of categories to problems

We invited the same two UX evaluators to assess both the quality of the ChatGPT-generated problem categories and the alignment between problem descriptions and their assigned categories. Having

previously analyzed the usability test videos, the evaluators were familiar with usability problems encountered by users. We first asked them to review the AI-generated problem categories. Both evaluators thought they were reasonable and comprehensive.

Recognizing that UX evaluators can employ different methods to categorize usability problems and that category labels can vary even when using the same method (Hornbæk & Frøkjær, 2008), we used AI-generated categories acknowledged by the evaluators to improve consistency among the evaluators. Then, we provided the evaluators with the list of AI-generated problem descriptions and asked them to assign the most appropriate category and subcategory to each problem separately. Following this, the evaluators discussed their categorizations to reach an agreement, resulting in a finalized list of problems and their assigned categories. We then compared the problem categories assigned by the evaluators with those assigned by ChatGPT, finding alignment rates of 86.7% for categories and 71.9% for subcategories.

Given that data categorization is inherently subjective and can exhibit bias even among human collaborators (Gao et al., 2024), AI-generated categories may also reflect biases stemming from the model's training data or linguistic tendencies. For example, we found that AI sometimes categorized usability problems based on their underlying causes. One instance is the problem "The user struggles to change the color of the text", which AI placed under "visibility of options and tools". However, the two human evaluators preferred to categorize it under "text editing issues", focusing on the specific function involved. To mitigate potential biases, *MultiUX* incorporates the definitions of AI-generated categories to improve users' understanding of AI outputs (as described in Section 4.1.3). It further enables users to review, supplement, merge, and edit AI-generated categories, thereby incorporating human judgment to calibrate potential AI biases.

5. User study

We conducted a between-subjects user study involving 20 participants to evaluate the usefulness and effectiveness of *MultiUX* and answer our research questions: **RQ1**-How does *MultiUX* support UX evaluators' performance in analyzing multiple usability test videos? **RQ2**-How does *MultiUX* influence UX evaluators' strategies for analyzing multiple videos? **RQ3**-What are UX evaluators' perceptions of how *MultiUX*'s key features support multiple video analysis?

5.1. Study design

To investigate how key features of *MultiUX* support multiple usability video analyses, such as categorizing problems and optimizing analysis sequence, we developed a baseline to isolate these key features by removing them. A screenshot of the baseline version is presented in Appendix C (Figure C1). The baseline interface lacked both a category panel (Figure 1B) and summary visualization (Figure 3A). Additionally, the videos were organized by participant number, without employing any recommendations based on problem similarity calculations. We chose to develop this version instead of leveraging an existing commercial platform to maintain a consistent UI between conditions and allow participants to focus on key features. In addition, the AI capability of our developed baseline was very close to the state-of-art tools, such as Lookback (Lookback, 2021) and UserTesting (UserTesting, 2021), which provided AI suggestions in the form of natural language. Because there are potential learning effects between conditions, we adopted a between-subjects design. For example, after a participant used *MultiUX*, they would know the categorization and recommendation features, which might prime them to consider this analytical strategy in the other condition.

5.2. Participants and apparatus

We recruited 20 participants (15 females, and 5 males) from design programs in local universities. Participants self-reported having 1–5 ($M = 2.7, SD = 0.9$) years of UX experience and were assigned to either *MultiUX* or baseline, with 10 participants each. The average years of UX experience for the two conditions were 2.5 and 2.8 years ($SD = 0.8, 1.0$) respectively. The median of self-reported familiarity

with usability analysis on a scale of 1–5 was the same for both conditions: "4 - very familiar". Mann-Whitney U tests found no significant differences in the years of UX experience or familiarity with usability analysis between the two conditions. Participants completed the study remotely, using their own computers to access our system and communicating with the moderator via video-conferencing software.

5.3. Study videos

We used the 10 processed usability test videos in Section 4.3 for the user study, with an average length of 1012s. Additionally, we collected another set of videos, featuring the same 10 users, which served as practice videos. In these videos, users were tasked with finding a suitable place for children to participate in cultural and creative activities on a Museum's website. The average length of these practice videos was 461 s.

5.4. Procedure

Each study session last approximately 120 min. Initially, participants watched a video tutorial covering the system's interface, including elements like the category panel and video recommendations. The moderator highlighted that AI suggestions may be inaccurate and participants hold the final responsibility for their analysis results. Following this, an explanation of the scenario and tasks for the usability videos was provided. Participants went through a training session to familiarize themselves with the system features, with the freedom to inquire until they felt confident. Subsequently, they proceeded to the formal usability video analysis, which had a set time limit, reflecting the typical time constraints faced by UX evaluators during usability video analyses (Fan et al., 2020b; Kuang et al., 2022). During the formal study, participants were asked to identify the usability problems of the product as completely as possible. Additionally, they needed to assess the frequencies of these identified problems. Participants were asked to complete a survey, which included a 5-point Likert Scale, to rate the effectiveness of the version they used after finishing the formal study. Finally, participants underwent a semi-structured interview covering their experience, feature usage, analytical strategies, and suggestions for system design. Participants were compensated for their time.

5.5. Data measurement and analysis

To evaluate user task performance and experience across the two conditions, we employed both quantitative and qualitative analyses. For objective measurements, we recorded user actions, video timestamps, and the problem list with associated categories. We tested for normality using the Shapiro-Wilk test. An independent t-test was conducted for normally distributed data, and the Mann-Whitney U test for non-normal data. Effect sizes were reported using Cohen's d (d) for normal data and rank-biserial correlation coefficient (r) for non-normal data. For subjective assessments, we recorded and transcribed the interview sessions and then analyzed them using the same method as in Section 3.

6. Results

This section presents the findings on the participants' multiple video analysis performance (**RQ1**), analytical strategies (**RQ2**), and perceptions of *MultiUX*'s features (**RQ3**). Participants using *MultiUX* and baseline are labeled M1 – 10 and B1-10, respectively.

6.1. Task performance

6.1.1. Video coverage

Participants viewed significantly more videos with *MultiUX* ($M_{MultiUX} = 8.5 > M_{baseline} = 3.1, p < 0.001$, $r = 0.88$) during the given time. The first reason was participants in the baseline took a longer time ($M_{baseline} = 11.9min > M_{MultiUX} = 1.7min, p < 0.001, r = 0.85$) to conduct bottom-up clustering of the

identified problems and derive the categories. Most of them (N = 7) switched videos several times to summarize the problem list when writing the report, noting down the frequencies of problems encountered by multiple users and avoiding missing problems. Although some participants tried to "count the frequencies during the analysis, given the multitude of problems, it was challenging to track frequency updates of problems until the study concluded."-B8 The visualization in *MultiUX* summarized problems across all viewed videos and displayed the categories agreed upon by participants, substantially reducing the time required for report writing. Another reason was without the AI-inferred categories as guidance, most participants (N = 9) using baseline chose to conduct detailed analyses of a limited number of videos to prioritize identifying more problems. B5 explained, "When I checked the AI suggestions, I noticed that the problem descriptions varied, indicating the emergence of new problems. I was concerned about potential oversights, which could lead to my evaluations not being as complete as intended."

6.1.2. Categories and frequencies of identified problems across multiple videos

At the end of the study, we asked participants to report product usability by putting similar usability problems into the same category and assessing the frequency of each category of problem (the number of users that encountered the problem). Participants identified significantly more categories ($M_{MultiUX} = 8.6 > M_{baseline} = 3, p < 0.001, r = 1$) and subcategories ($M_{MultiUX} = 20.3 > M_{baseline} = 9.9$, $p < 0.001, d = 2.98$) in *MultiUX*, showing that *MultiUX* could support identifying a broader range of usability problems. Figure 5 illustrates the frequency distribution of the identified problem categories, demonstrating that participants using *MultiUX* identified more categories with higher frequencies. For subcategories, problems encountered by 1–3 users accounted for 75.5% and 98%, 4–6 users for 22.5% and 1%, and 7–9 users for 2% and 1% respectively in *MultiUX* and baseline. Compared with the categories, the frequencies of subcategories were slightly lower, which was reasonable considering the fine granularity and the wide variety of problems each user encountered.

6.2. Participants' strategies for analyzing multiple videos

Most participants (N = 9) in the baseline chose to conduct detailed analysis to guarantee the problem completeness, finding no effective strategies to analyze all videos and compare them to assess the problem frequencies. On average, they analyzed two or three videos before starting to summarize and write the report. In contrast, all participants (N = 10) using *MultiUX* first analyzed two or three videos in detail to check the accuracy of AI suggestions and add problems overlooked by AI, aiming to identify a wide range of problems (Figure 6,a1). Then, they followed the guidance of categories to analyze the remaining videos and assess the problem frequencies quickly (Figure 6,a2). Referring to the problem distribution visualization, they wrote the usability report involving all problems with frequencies organized by categories efficiently. Participants in both conditions shared some similar analytical behaviors when identifying and recording problems, such as "Pause-Write" regarding the confirmed problems and "Playback-Write" to re-check and evaluate the problems. "Fast-check" means jumping and viewing the highlighted timestamps quickly. *MultiUX* enabled participants to leverage this strategy to assess the frequencies efficiently (Figure 6,a2), while baseline participants exhibited this behavior in two stages:

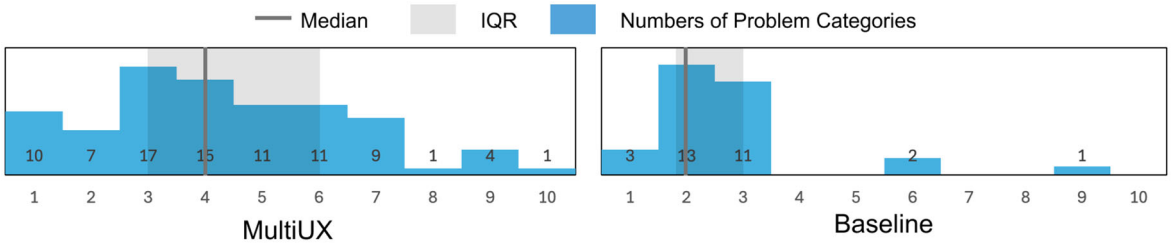

Figure 5. Bar chart with the x-axis showing how many videos featured the same problem category, and the y-axis representing the number of problem categories. We compared two conditions: *MultiUX* and baseline. Since each UX evaluator identified a different number of problem categories, the y-axis totals differ between the two conditions. Overall, participants using *MultiUX* identified a greater number of problem categories, with problems within the same category found in more videos.

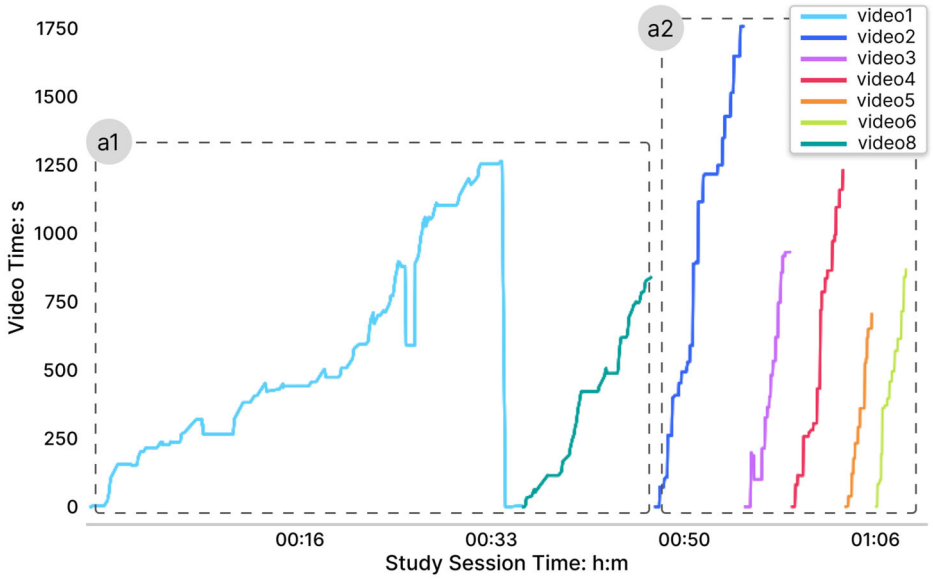

Figure 6. Analytical behaviors of a *MultiUX* participant across multiple videos (x-axis: study session time, y-axis: video time viewed): (a1) detailed analysis, where the participant reviewed the video from start to finish, pausing and replaying to identify comprehensive usability problems; and (a2) fast-check, where the participant followed the guidance of AI-generated categories to jump between timestamps, efficiently assessing problem frequencies. This participant was selected because his behavior exemplifies the primary strategies for analyzing multiple videos.

reviewing the analytical output of a single video and summarizing problems across all viewed videos to write a usability report.

6.2.1. Determine analytical sequence

Most participants (N = 9) in the baseline chose to analyze videos in sequence or randomly selected one to start, except for B4, "I chose to start from the one with the most AI suggestions, as I believed it indicated the user encountering the most problems." Among participants using *MultiUX*, four selected the first video randomly, and the other six employed two strategies to guide the initial selection: (1) video with the most categories (N = 5), and (2) video with the most "User interface and experience issues" (N = 1). Participants believed that the video with the most categories could help them identify a broader range of usability problems. M4 initially selected the video with the most "User interface and experience issues" to focus on interaction difficulties caused by the product design, rather than the issues about the AI capabilities of the test product. After finishing the detailed analysis of the first selected video, all participants (N = 10) chose the least similar video from the recommendation list to identify the most diverse set of problems. Participants then started to assess the frequencies of these identified problems.

6.2.2. Assess frequencies of identified problems

Most participants (N = 7) in the baseline assessed frequencies by relying on their memory and usability problem lists. B2 said, "I initially wrote down the more prominent problems and categorized them based on memory. Subsequently, I checked the problem lists to identify any missing problems and placed them under the appropriate categories." They jumped back to the source videos to re-check these identified problems quickly and categorize them more confidently. Conversely, B3 and B6 wrote the report solely relying on memory, "I was recording these problems just now, so, I believed the accuracy of my memory to be around 90%." Additionally, B1 copied all the problems to Figma [2] to conduct a bottom-up clustering.

To assess the frequencies efficiently, *MultiUX* guided all participants (N = 10) with categories to confirm whether any other users encountered similar problems in a limited time. M1 and M2 assessed the frequencies by viewing each problem category individually, employing the "Recommendations for problems within the same category" (Figure 2C) and "Usability Problem Category" (Figure 1B) filter

features. They transitioned between different videos numerous times. "To avoid the slowdown in assessment speed caused by the frequent transition"-M5, the other eight participants analyzed the videos one by one, focusing on whether problems identified previously were present in these remaining videos. In addition to assessing frequencies among multiple users, participants noticed that some users encountered the same problem many times and identified these recurring problems using the categories. To speed up the problem identification and frequency assessment, some participants (N = 4) shifted their strategies from reviewing the video segments to reading the available text. They cross-referenced problem descriptions and categories, and AI explanations in the text format to determine whether the AI suggestions were indeed relevant problems.

6.3. Perceptions of how MultiUX support multiple video analyses

The post-study survey (5-point scale) demonstrated that participants thought both conditions were **(1) easy to use:** *MultiUX* ($Md = 4, IQR = 1.75$), and baseline ($Md = 4, IQR = 0$); **(2) satisfactory:** *MultiUX* ($Md = 4, IQR = 0$), and baseline ($Md = 4, IQR = 0.75$); and **(3) trustworthy:** *MultiUX* ($Md = 4, IQR = 0.75$) and baseline ($Md = 4, IQR = 0$). Although with higher learning cost ($Md_{MultiUX} = 3 > Md_{baseline} = 1$), *MultiUX* could save time for analyzing multiple videos ($Md_{MultiUX} = 4 > Md_{baseline} = 3.5, p < 0.01, r = .7$) and support assessing the problem frequencies across multiple videos ($Md_{MultiUX} = 5 > Md_{baseline} = 2, p < 0.001, r = 1$).

6.3.1. Problem categorization for video overview and problem recommendation

Participants used problem categorizations on the video timeline in two main ways: (1) act as an overview of all videos before conducting a detailed analysis; and (2) act as a guide to assess the generalization of problems, determining whether they were commonly encountered by multiple users or specific to individual characteristics.

Most participants (N = 6) agreed that the AI-assigned categories were reasonable and accurate. These categories facilitated the assignment of videos among colleagues, which is a common practice for collaborative usability analysis (Kuang et al., 2022). M1 noted, *I could efficiently review the problem categories encountered by each user and assign videos based on evaluators' familiarity with the categories."* Additionally, M7 highlighted the benefit of reviewing problems within the same category to determine if they were truly usability problems: "sometimes problems arise due to a specific user's habits. Other users do not encounter similar problems, so we do not consider them real problems." The accuracy of AI-assigned categories also increased participants' willingness to use other AI-assisted features, such as AI-recommended categories for manually added problems (M8). However, M2 pointed out a limitation of the AI, "for similar problems across different videos, AI occasionally assigned different categories because users verbalized them differently."

When manually adding usability problems, seven participants either selected existing categories or created new ones directly. "Since the problems were identified by myself, I was very confident in their categorizations."-M7. Some participants preferred using the AI recommendation feature because "the recommended ones were quite accurate, totally aligning with the problem descriptions."-M2. M8 shared her strategy to get more accurate recommendations, "I emulated the logic and language used by AI when describing problems and provided highly detailed descriptions."

6.3.2. Category system for problem management and filtering

All participants using *MultiUX* felt that the category panel supported them in identifying more complete usability problems, identifying problems within the same category, and efficiently assessing problem frequencies across multiple videos. M6 further emphasized that the categories could streamline the subsequent assignment of usability problems. For example, "functionality problems" should be assigned to the development team, while "user interface and experience problems" should be directed to the design team to resolve.

The AI-inferred categories were praised as "highly complete" (M8) and "saved UX evaluators' time and effort in considering and defining them" (M1, M4). However, feedback on the AI-inferred subcategories varied. M1, M2, and M3 found them too detailed, leading to overlap and requiring further

review and refinement by UX evaluators. In contrast, M4 appreciated their level of granularity, noting that they helped pinpoint specific usability problems, such as "text editing problems." Some participants (N = 4) took additional time learning and familiarizing the names and definitions of these AI-inferred categories before or during their analysis, potentially increasing the learning cost.

Some participants (N = 4) noticed similarities among categories after adding new ones. Thus, they employed the category optimization feature to optimize the existing category set. Participants noted that *MultiUX* now provided optimization suggestions, such as merging two categories, mainly through "analyzing the similarity based on categories' names and definitions"-M6. It would be better if "AI could further analyze the reasons for each category of problems and suggest merging the categories with different definitions but caused by the same reason"-M3.

6.3.3. Video recommendations to optimize multiple video analysis sequences

Section 6.2.1 described how video recommendations helped participants determine the analysis sequence across multiple videos. M1 appreciated this feature, noting that "previously, problem similarities between users were only summarized after analyzing all videos. MultiUX's preliminary analyses allowed for a more strategic and focused approach when analyzing multiple videos." M3 valued the real-time adjustments in recommendations based on her selections, stating, "the recommendation list updated after I chose a category, which was valuable for focused analysis. For instance, it might help me understand why some users faced 'functionality issues' while others did not."

All participants (N = 10) selected the least similar video as their second target of analysis in the study, which helped them "identify more categories of problems that had not appeared before, aiming to enhance the problem completeness" - M5. This strategy of alternating between similar and dissimilar videos led to a more comprehensive analysis within the study's time constraints. Acknowledging that time limitations might influence participants' usage of video recommendations, we asked them to reflect on how this feature could be applied in real-world practices, based on their experience in the study and daily work routines.

Participants noted that they would also use video recommendations to identify common problems, facilitating severity assessments and problem-fix prioritizations. This is similar to their usage of problem recommendations and category panels. After analyzing multiple videos, M6 described the recommendations as "accurate" and "reasonable," appreciating how it assisted in recognizing cross-video comparisons and problem recurrence. M10 highlighted, "it was helpful to see how similar problems occurred in different contexts and deepened my understanding of the problem severity." M9 reflected that focusing on similar videos allowed UX evaluators to make better decisions regarding redesign priorities.

6.3.4. Visualization of problem distribution by categories across multiple videos

Summary visualization was used throughout the whole analysis process, including (1) Guiding the selection of the first video for analysis (section 6.2.1); (2) Providing the overview of the problems encountered by users during the process of analysis; (3) Acting as a reference for report writing. M2 elaborated on a specific usage scenario: "I quickly identified through visualization which users encountered a category of less common problems, and then proceeded to review the corresponding source videos." M9 described its usefulness for report writing, "I found the visualization to be highly intuitive. It enabled me to view both which users encountered the same category of problems and the different problem categories experienced by each user."

7. Discussion

7.1. Empowering multiple video analysis with MultiUX

MultiUX addresses the challenge of determining the optimal analysis sequence when faced with a large volume of videos and limited time. This challenge also exists in other domains of multi-video analysis, such as analyzing sports training videos (Wang et al., 2019). Our findings revealed the benefits of guiding user navigation in multiple videos, which could inspire similar navigational designs in other domains.

Unlike previous studies that focus primarily on segmenting multiple videos based on object or gesture detection (Choudhury et al., 2020; Poleg et al., 2014), *MultiUX* was developed as a platform that not only segments videos but also categorizes the clips. A comparable tool is designed to analyze group activity using multiple egocentric videos (Charoenkulvanich et al., 2019), which also includes a filtering feature by categories across multiple videos. Their results demonstrate that this approach supports users to efficiently review relevant clips in different videos, aligning with our observations, and reinforces the effectiveness of the category system for multiple video analyses.

Furthermore, *MultiUX* presents an interactive sunburst view that offers both an overview of the problems in all videos and summaries in each video. This design is more extensive than common time series visualizations, which are generally limited to single video analyzes (Li et al., 2010). Furthermore, due to its intuitive design, it addresses the drawbacks of time series visualizations, which can provide users with too much information, leading to distraction and confusion (Charoenkulvanich et al., 2019). The sunburst view demonstrates potential for wider use in other domains of multiple video analyses.

7.2. Positioning MultiUX among existing usability analysis tools

To better understand where *MultiUX* fits within the broader landscape of usability analysis tools, we compared its design and functionality with both commercial and research tools for analyzing usability test videos. Following a recent survey on tools commonly used by UX professionals, we selected five representative commercial tools, including Maze[3], UserTesting[4], Dovetail[5], Lookback[6], and Optimal Workshop[7] (Tools, 2025). The most comparable research system is "UX Assistant" proposed by Kuang et al (Kuang et al., 2024).

All these tools integrate AI to assist in analyzing transcripts of usability test sessions. For example, UserTesting automatically labels transcripts with tags such as "like", "confusion", and "suggestion". Maze, Dovetail, Lookback, Optimal Workshop, and UX Assistant employ AI for semantic understanding of users' feedback, generating summaries or inferring user challenges. Similarly, *MultiUX* uses AI to identify potential usability problems from the transcripts of usability test videos.

Synthesizing feedback from multiple users to identify common problems has recently become a prominent feature in these tools, and it is also a central concept in the design of *MultiUX*. Tools such as Maze, UserTesting, Dovetail, and Optimal Workshop now include "theme generation" or "AI insight summary" features that automatically highlight recurring challenges across users. However, these tools emphasize speed and automation, and the process by which insights are synthesized remains largely opaque. Although some tools link generated insights back to source clips as supporting evidence, human evaluators have limited visibility into how those insights are derived.

In contrast, *MultiUX* emphasizes process transparency and human control. It supports human evaluators in identifying, comparing, and categorizing usability problems across multiple videos, guiding them to derive final insights through structured analysis. This workflow aligns with evaluators' natural practices and clearly show how each insight is developed. Furthermore, evaluators can edit AI outputs at every stage, making *MultiUX* a genuinely human-AI collaborative system rather than a purely automated one. In addition, the "video recommendation" feature in *MultiUX* helps evaluators prioritize which videos to analyze next, a need reported by evaluators working under time constraints.

However, mature commercial tools offer broader functionality beyond analysis. They support participant recruitment, testing execution, and integration with various data sources. For example, Maze and UserTesting capture users' interaction paths, while Dovetail integrates data from bug reports and customer calls. Such additional data can provide richer evidence for identifying usability problems. In comparison, *MultiUX* focuses narrowly but deeply on cross-video problem categorization. Its emphasis on transparency and human-AI collaboration offers design implications for commercial tools, encouraging them to make their analytic processes more interpretable and to foster human expert involvement. In the future, *MultiUX* could similarly benefit from incorporating diverse data sources beyond usability test videos, thereby providing a more comprehensive foundation for identifying and prioritizing usability problems.

7.3. Novelty of MultiUX's human-AI collaboration

MultiUX enhances multiple usability video analyses by integrating AI to collaborate with UX evaluators throughout the process. The AI provides real-time interactions based on users' actions, such as video and problem recommendations and category optimizations, improving the efficiency and confidence of the analysis. Participants noted that the AI's real-time responses were accurate, and the collaboration was flexible, supporting their decision-making in determining the analysis sequence, as well as in comparing and categorizing problems during the analysis.

In contrast, previous studies have used AI primarily for video segmentation in multiple video analysis (Charoenkulvanich et al., 2019; Panda et al., 2017). In these workflows, AI segments the videos, and humans compare and categorize these segments for further analysis. However, there is no real-time interaction or opportunity to modify AI-generated output. The continuous interaction between humans and AI in *MultiUX* can be applied to other domains of multiple video analysis, such as sports training (Xia et al., 2014) and group activity analysis (Charoenkulvanich et al., 2019) to increase the level of AI support.

Furthermore, unlike previous studies where AI only segmented videos during offline pre-processing, *MultiUX* goes a step further to compare and categorize these segments to provide initial categories as references. This approach saves time in organizing and defining categories and could also be applied to other domains to improve the efficiency of comparison and categorization. For instance, a previous study used keyword search techniques to identify goal events in sports games videos based on web-casting text (Zhu et al., 2007). In a similar way, ChatGPT could be leveraged to further compare and categorize these texts with the same keywords, inferring the opinions about the goal event.

7.4. Design implications for future tool design

Participants in our study demonstrated varying preferences for the level of AI automation. Some preferred AI to proactively recommend analysis sequences and categorize problems across multiple videos, while others wanted more control, making these decisions independently. These differences may reflect individual expertise and work habits (Borlinghaus & Huber, 2021; Hertzum & Jacobsen, 2001; Khajouei et al., 2011). Although *MultiUX* allows adjustments to automation levels, participants tended to overlook it and use the pre-defined categories, only adding new ones to enhance completeness. Therefore, it is essential to prioritize **the customization of AI automation levels** at the beginning of the analysis process. For example, upon entering *MultiUX*, UX evaluators should select level of automation from three options: (1) no AI assistance, (2) AI assistance upon request, or (3) full AI assistance (Mackeprang et al., 2019). The selected level would decide when AI-generated data is presented.

Second, we addressed the need for **flexibility in the categories** to manage problems across multiple videos. Most of the participants in the formative study preferred self-defined categories over standard ones such as Nielsen's heuristics, which led us to integrate AI-generated categories. However, during the study, five participants expressed the desire for additional categorization options. For example, B10 stated, "The AI-generated categories worked well to group similar problems in multiple videos, but I added severity ratings to prioritize certain problems." This highlights the importance of offering diverse categorization options to meet varying needs. The participants also noted the subjectivity and uncertainty in categorizing usability problems across videos. M4 reflected that "I questioned whether there could be a universal set of categories for all products, yet I was not sure how to establish such a standard." While *MultiUX* may not provide a one-size-fits-all set of categories, offering multiple category sets would allow flexible combinations, better supporting the diverse needs when analyzing usability problems in multiple videos.

Third, participants expressed the need for the participation of colleagues in certain instances, which is common in usability analysis (Soure et al., 2022). For example, M10 mentioned, "When unsure about AI suggestions, I prefer to discuss with colleagues before deciding whether to use them." Future designs should incorporate **collaboration scenarios** that support synchronous and asynchronous analysis by multiple UX evaluators. In addition, it should accommodate various modes of collaboration. For example, M1 noted that under time constraints, they often distributed videos among colleagues, which required awareness of the progress of each collaborator. In another scenario, to enhance the reliability and completeness of the analysis, multiple evaluators needed to collaboratively analyze the same videos

(Soure et al., 2022). Future designs should support these modes by tracking collaborators' progress and providing a discussion panel to resolve disagreements.

Finally, while AI explanations in *MultiUX* have helped UX evaluators understand the reasoning behind problem categorizations and video recommendations, an **interactive conversational AI assistant** could offer an additional way to seek clarification and gain deeper insights from AI. Although our participants did not specifically request an AI assistant, previous studies suggest that such an assistant could benefit usability analysis by having knowledge of all videos and could compare them to identify common problems (Kuang et al., 2023). Thus, incorporating an AI assistant as an additional interaction paradigm for visual analytics interfaces could boost the support provided to UX evaluators and enhance multiple video analyses.

7.5. Limitations and future work

Our research has effectively designed a human-AI collaborative tool *MultiUX*, to enhance the analysis of multiple usability videos by optimizing the analysis sequences and categorizing the identified problems. However, we acknowledge several limitations associated with the controlled lab study setting.

First, *MultiUX* was tested with participants recruited from design programs. Although many had internship experience, their expertise and analytical strategies could differ slightly from those of experienced UX evaluators. For instance, prior studies suggest that novice evaluators benefit from structured formats for usability problem annotation, whereas experienced evaluators typically do not (Følstad et al., 2012a). Future research should involve a more diverse group of participants to explore whether *MultiUX* has diverse influences on different groups of UX evaluators.

Secondly, although participants were instructed to analyze as they would in professional settings, the study design and time constraints could have influenced their analytical behavior. For example, when AI suggestions slightly deviated from their work habits, participants often chose not to make modifications to save time. This contrasts with their work practice, where they would provide detailed and accurate descriptions to facilitate discussions with colleagues. Similarly, for uncertain categories, participants noted that in actual projects, they would consult with colleagues before making a decision. However, they opted for the category that they felt was most appropriate during the study. A longitudinal study would provide deeper insights into how UX evaluators customize AI-inferred categories and how video recommendation usage evolves over longer product development cycles.

Lastly, although we tested *MultiUX* on datasets from different product types (website, mobile application, and VR application) to examine its domain generality, our user study involving UX evaluators focused on multiple usability videos of website products. This choice was made to ensure consistency in task complexity and to maintain a manageable study duration. Future work could extend the user study to include usability videos of other product types to further validate the effectiveness of *MultiUX*.

8. Conclusion

We took the first step to explore how human-AI collaboration supports analyzing multiple usability test videos. Drawing from insights in a formative study, we developed *MultiUX*, a human-AI collaborative tool that integrates AI-categorized usability problems, enabling the management and recommendation of problems across multiple videos. Additionally, *MultiUX* recommends videos based on problem similarities to optimize analysis sequences and includes a summary visualization to display problem distribution by categories across all videos. We conducted a between-subjects study to examine how *MultiUX* and its key features support multiple video analysis. We developed a baseline version without these features to isolate their impact. Results showed that participants using *MultiUX* analyzed more videos within the given time, identified a broader range of problems, and reported higher problem frequencies compared to the baseline. Furthermore, participants more strategically determined analysis sequences and assessed problem frequencies with *MultiUX*'s support. They also noted that the key features addressed challenges in multiple video analyses and enhanced their workflow. Building upon these findings, we proposed several design implications to enhance future tool development for supporting multiple usability test video analyses.

Notes

1. Hereafter interchangeably referred to as "AI"
2. https://www.figma.com/
3. https://maze.co/
4. https://www.usertesting.com/platform-overview
5. https://dovetail.com/
6. https://www.lookback.com/
7. https://www.optimalworkshop.com/

Author contributions

CRediT: **Luyao Shen**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft, Writing – review & editing; **Qing Shi**: Software; **Emily Kuang**: Conceptualization, Writing – review & editing; **Linjie Qiu**: Software; **Shixu Zhou**: Software; **Pan Hui**: Funding acquisition, Supervision, Writing – review & editing; **Mingming Fan**: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Disclosure statement

No potential competing interest was reported by the author(s).

Funding

This work is partially supported by the Guangzhou Municipal Nansha District Science and Technology Bureau under Contract No.2022ZD012, Guangzhou-HKUST(GZ) Joint Funding Project (No.: 2024A03J0617), Education Bureau of Guangzhou Municipality Funding Project (No.: 2024YBJG070), The Guangzhou Science and Technology Program City-University Joint Funding Project (No.: 2023A03J0001), Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.: 2023B1212010007), and AI Research and Learning Base of Urban Culture under Project (No.: 2023WZJD008).

ORCID

Luyao Shen http://orcid.org/0009-0006-8403-6271
Emily Kuang http://orcid.org/0000-0003-4635-0703
Mingming Fan http://orcid.org/0000-0002-0356-4712

References

Batch, A., Ji, Y., Fan, M., Zhao, J., & Elmqvist, N. (2024). uxsense: Supporting user experience analysis with visualization and computer vision. *IEEE Transactions on Visualization and Computer Graphics*, *30*(7), 3841–3856. https://doi.org/10.1109/TVCG.2023.3241581

Borlinghaus, P., & Huber, S. (2021). *Comparing apples and oranges: Human and computer clustered affinity diagrams under the microscope* [Paper presentation]. Proc. ACM IUI, IUI '21, New York, NY, USA (pp. 413–422). Association for Computing Machinery. https://doi.org/10.1145/3397481.3450674

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Burnam, L. (2023). We surveyed 1093 researchers about how they use ai-here's what we learned.

Charoenkulvanich, N., Kamikubo, R., Yonetani, R., & Sato, Y. (2019). *Assisting group activity analysis through hand detection and identification in multiple egocentric videos* [Paper presentation]. Proc. ACM IUI, IUI '19, New York, NY, USA (pp. 570–574). Association for Computing Machinery. https://doi.org/10.1145/3301275.3302297

Chen, Y., Li, K., Bao, W., Patel, D., Kong, Y., Min, M. R., & Metaxas, D. N. (2025). Learning to localize actions in instructional videos with llm-based multi-pathway text-video alignment. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024* (pp. 193–210). Springer Nature Switzerland.

Choudhury, A. M., Saif, A. F. M. S., & Rahman, M. (2020). *Toddler sensory-motor development for object manipulation by analyzing hand-pose* [Paper presentation]. Proc. ICCA, ICCA 2020, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3377049.3377055

Dai, S.-C., Xiong, A., & Ku, L.-W. (2023). Llm-in-the-loop: Leveraging large language model for thematic analysis. *arXiv*.

De Paoli, S. (2023). Improved prompting and process for writing user personas with llms, using qualitative interviews: Capturing behaviour and personality traits of users. *arXiv*.

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

Delice, E. K., & Güngör, Z. (2009). The usability analysis with heuristic evaluation and analytic hierarchy process. *International Journal of Industrial Ergonomics*, *39*(6), 934–939. https://doi.org/10.1016/j.ergon.2009.08.005

Dunivin, Z. O. (2025). Scaling hermeneutics: A guide to qualitative coding with llms for reflexive content analysis. *EPJ Data Science*, *14*(1), 28. https://doi.org/10.1140/epjds/s13688-025-00548-8

Fan, M., Li, Y., & Truong, K. N. (2020a). Automatic detection of usability problem encounters in think-aloud sessions. *ACM Transactions on Interactive Intelligent Systems*, *10*(2), 1–24.). https://doi.org/10.1145/3385732

Fan, M., Lin, J., Chung, C., & Truong, K. N. (2019). Concurrent think-aloud verbalizations and usability problems. *ACM Transactions on Computer-Human Interaction*, *26*(5), 1–35.). https://doi.org/10.1145/3325281

Fan, M., Shi, S., & Truong, K. N. (2020b). Practices and challenges of using think-aloud protocols in industry: An international survey. *J. Usability Stud*, *15*(2), 85–102. https://dl.acm.org/doi/10.5555/3532708.3532711

Fan, M., Wu, K., Zhao, J., Li, Y., Wei, W., & Truong, K. N. (2020c). Vista: Integrating machine intelligence with visualization to support the investigation of think-aloud sessions. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 343–352. https://doi.org/10.1109/TVCG.2019.2934797

Fan, M., Yang, X., Yu, T., Liao, Q. V., & Zhao, J. (2022). Human-ai collaboration for ux evaluation: Effects of explanation and synchronization. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–32. https://doi.org/10.1145/3512943

Fan, M., Zhao, Q., & Tibdewal, V. (2021). *Older adults' think-aloud verbalizations and speech features for identifying user experience problems* [Paper presentation]. Proc. ACM CHI, CHI '21, New York, NY, USA. In Association for Computing Machinery. https://doi.org/10.1145/3411764.3445680

Fan, M., & Zhu, L. (2021). *Think-aloud verbalizations for identifying user experience problems: Effects of language proficiency with chinese non-native english speakers* [Paper presentation]. Proc. ACM ChineseCHI, Chinese CHI '21, New York, NY, USA (pp. 22–32). Association for Computing Machinery. https://doi.org/10.1145/3490355.3490358

Følstad, A., Law, E., & Hornbæk, K. (2012a). *Analysis in practical usability evaluation: A survey study* [Paper presentation]. Proc. ACM CHI, CHI '12, New York, NY, USA (pp. 2127–2136). Association for Computing Machinery.

Følstad, A., Law, E. L.-C., & Hornbæk, K. (2012b). *Outliers in usability testing: How to treat usability problems found for only one test participant?* [Paper presentation]. Proc. ACM NordiCHI, NordiCHI '12, New York, NY, USA (pp. 257–260). Association for Computing Machinery.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. (2025). Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 24108–24118).

Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., & Perrault, S. T. (2024). Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. *arXiv*. https://doi.org/10.1145/3613904.3642002

Gebreegziabher, S. A., Zhang, Z., Tang, X., Meng, Y., Glassman, E. L., & Li, T. J.-J. (2023). *Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis* [Paper presentation]. Proc. ACM CHI, CHI '23, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581352

Georgsson, M., & Staggers, N. (2016). An evaluation of patients' experienced usability of a diabetes mhealth system using a multi-method approach. *Journal of Biomedical Informatics*, *59*(C), 115–129. https://doi.org/10.1016/j.jbi.2015.11.008

Ghosh, S., Malempati, K., & Charette, C. (2025). Human-ai collaborative content analysis: Investigating the efficacy and challenges of llm-assisted content analysis for tiktok videos on palliative care. *Proceedings of the Association for Information Science and Technology*, *62*(1), 229–240. https://doi.org/10.1002/pra2.1251

Hagedorn, J., Hailpern, J., & Karahalios, K. G. (2008). *Vcode and vdata: Illustrating a new framework for supporting the video annotation workflow* [Paper presentation]. Proc. ACM AVI, AVI '08, New York, NY, USA (pp. 317–321) Association for Computing Machinery.

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, *13*(4), 421–443. https://doi.org/10.1207/S15327590IJHC1304_05

Hornbæk, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers*, *20*(6), 505–514. https://doi.org/10.1016/j.intcom.2008.08.005

Hwang, W., & Salvendy, G. (2010). Number of people required for usability evaluation: The 10 ± 2 rule. *Communications of the ACM*, *53*(5), 130–133. https://doi.org/10.1145/1735223.1735255

Jeddi, F. R., Nabovati, E., Bigham, R., & Farrahi, R. (2020). Usability evaluation of a comprehensive national health information system: A heuristic evaluation. *Inform. Med. Unlocked*, *19*, 100332. https://doi.org/10.1016/j.imu.2020.100332

Kang, D., Han, Z., Tian, J., Zhang, M., & Rzeszotarski, J. M. (2025). Themeviz: Understanding the effect of human-ai collaboration in theme development with an llm-enhanced interactive visual system. *Proceedings of the ACM on Human-Computer Interaction*, *9*(7), 1–29. https://doi.org/10.1145/3757675

Keenan, S. L., Hartson, H. R., Kafura, D. G., & Schulman, R. S. (1999). The usability problem taxonomy: A framework for classification and analysis. *Empirical Software Engineering*, *4*(1), 71–104. https://doi.org/10.1023/A:1009855231530

Khajouei, R., Peute, L. W., Hasman, A., & Jaspers, M. W. (2011). Classification and prioritization of usability problems using an augmented classification scheme. *Journal of Biomedical Informatics*, *44*(6), 948–957. https://doi.org/10.1016/j.jbi.2011.07.002

Kjeldskov, J., Skov, M. B., & Stage, J. (2004). *Instant data analysis: Conducting usability evaluations in a day* [Paper presentation]. Proc. ACM NordiCHI, NordiCHI '04, New York, NY, USA (pp. 233–240). Association for Computing Machinery.

Kocaballi, A. B. (2023). Conversational ai-powered design: Chatgpt as designer, user, and product. *arXiv*.

Kuang, E., Jahangirzadeh Soure, E., Fan, M., Zhao, J., & Shinohara, K. (2023). *Collaboration with conversational ai assistants for ux evaluation: Questions and how to ask them (voice vs. text)* [Paper presentation]. Proc. ACM CHI, CHI '23, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581247

Kuang, E., Jin, X., & Fan, M. (2022). *"merging results is no easy task": An international survey study of collaborative data analysis practices among ux practitioners* [Paper presentation]. Proc. ACM CHI, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3491102.3517647

Kuang, E., Li, M., Fan, M., & Shinohara, K. (2024). *Enhancing ux evaluation through collaboration with conversational ai assistants: Effects of proactive dialogue and timing* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA. Association for Computing Machinery.

Li, H., Hao, Y., Zhai, Y., & Qian, Z. (2023). *Assisting static analysis with large language models: A chatgpt experiment* [Paper presentation]. Proc. ACM FSE, ESEC/FSE 2023, New York, NY, USA (pp. 2107–2111). Association for Computing Machinery.

Li, H., Tang, J., Wu, S., Zhang, Y., & Lin, S. (2010). Automatic detection and analysis of player action in moving background sports video sequences. *IEEE Trans. Circuits Syst*, *20*(3), 351–364. https://doi.org/10.1109/TCSVT.2009.2035833

Liew, J., McCracken, N., & Crowston, K. (2014a). *Semi-automatic content analysis of qualitative data* [Paper presentation]. Proc. IConference.

Liew, J., McCracken, N., Zhou, S., & Crowston, K. (2014b). Optimizing features in active machine learning for complex qualitative content analysis. In *Proc. ACL* (pp. 44–48).

Liljegren, E. (2006). Usability in a medical technology context assessment of methods for usability evaluation of medical equipment. *International Journal of Industrial Ergonomics*, *36*(4), 345–352. https://doi.org/10.1016/j.ergon.2005.10.004

Lookback. (2021). *Lookback: Simple and powerful user research*. https://lookback.io/.

Lookback (2024). Lookback.

Lu, Y., Yang, Y., Zhao, Q., Zhang, C., & Li, T. J.-J. (2024). Ai assistance for ux: A literature review through human-centered AI. *arXiv*.

Mackeprang, M., Müller-Birn, C., & Stauss, M. T. (2019). Discovering the sweet spot of human-computer configurations: A case study in information extraction. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30.). https://doi.org/10.1145/3359297

Marathe, M., & Toyama, K. (2018). *Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes* [Paper presentation]. Proc. ACM CHI, New York, NY, USA (pp. 1–12) Association for Computing Machinery.

Marcio Silva, C., Macedo, V., Lemos, R., & Okimoto, M. L. L. (2014). Evaluating quality and usability of the user interface: A practical study on comparing methods with and without users. In *Proc. DUXU* (pp. 318–328). Springer Springer.

McDonald, S., Edwards, H. M., & Zhao, T. (2012). Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication*, *55*(1), 2–19. https://doi.org/10.1109/TPC.2011.2182569

Mugunthan, T. (2023). Researching the usability of early generative-ai tools.

Nielsen, J. (2012). How many test users in a usability study?.

Nielsen, J. (2022). Severity ratings for usability problems: Article by jakob nielsen.

Nielsen, J. (2023). Why you only need to test with 5 users.

Nielsen, J. (2024). 10 usability heuristics for user interface design.

Nørgaard, M., & Hornbæk, K. (2006). *What do usability evaluators do in practice? an explorative study of think-aloud testing* [Paper presentation]. Proc. ACM DIS, New York, NY, USA (pp. 209–218) Association for Computing Machinery.

OpenAI. (2024a). Models - openai api.

OpenAI. (2024b). Prompt engineering.

Overney, C., Saldías, B., Dimitrakopoulou, D., & Roy, D. (2024). *Sensemate: An accessible and beginner-friendly human-ai platform for qualitative data analysis* [Paper presentation]. Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI '24, New York, NY, USA (pp. 922–939). Association for Computing Machinery.

Panda, R., Mithun, N. C., & Roy-Chowdhury, A. K. (2017). Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, *26*(10), 4712–4724. https://doi.org/10.1109/TIP.2017.2708902

Petrie, H., & Power, C. (2012). *What do users really care about? a comparison of usability problems found by users and experts on highly interactive websites* [Paper presentation]. Proc. ACM CHI, New York, NY, USA (pp. 2107–2116) Association for Computing Machinery.

Poleg, Y., Arora, C., & Peleg, S. (2014). *Temporal segmentation of egocentric videos* [Paper presentation]. Proc. IEEE CVPR (pp. 2537–2544).

Qiao, T., Walker, C., Cunningham, C., & Koh, Y. S. (2025). *Thematic-lm: A llm-based multi-agent system for large-scale thematic analysis* [Paper presentation]. Proceedings of the ACM on Web Conference 2025, WWW '25, New York, NY, USA (pp. 649–658) Association for Computing Machinery.

Rietz, T., & Maedche, A. (2020). Towards the design of an interactive machine learning system for qualitative coding. In *Proc. ICIS*.

Rietz, T., & Maedche, A. (2021). *Cody: An ai-based system to semi-automate coding for qualitative research* [Paper presentation]. Proc. ACM CHI, CHI '21, New York, NY, USA (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445591

Ronanki, K., Cabrero-Daniel, B., & Berger, C. (2022). Chatgpt as a tool for user story quality evaluation: Trustworthy out of the box?. In *Proc. XP* (pp. 173–181). Springer.

Rosala, M. (2024). How to analyze qualitative data from ux research: Thematic analysis.

Sauro, J. (2014). The relationship between problem frequency and problem severity in usability evaluations. *J. Usability Stud*, *10*(1), 17–25. https://dl.acm.org/doi/10.5555/2817310.2817312

Schiavone, W., Roberts, C., Du, D., Sauro, J., & Lewis, J. (2023). Can chatgpt replace ux researchers? an empirical analysis of comment classifications.

Singh, S. H., Jiang, K., Bhasin, K., Sabharwal, A., Moukaddam, N., & Patel, A. B. (2024). Racer: An llm-powered methodology for scalable analysis of semi-structured mental health interviews. *arXiv*

Soure, E. J., Kuang, E., Fan, M., & Zhao, J. (2022). Coux: Collaborative visual analysis of think-aloud usability test videos for digital interfaces. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 643–653. https://doi.org/10.1109/TVCG.2021.3114822

Stasko, J., & Zhang, E. (2000). Focus + context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings* (pp. 57–65).

Sun, A., Bhowmick, S. S., & Chong, J.-A. (2011). *Social image tag recommendation by concept matching* [Paper presentation]. Proc. ACM Int. Conf. Multimed., MM '11, New York, NY, USA (pp. 1181–1184) Association for Computing Machinery. https://doi.org/10.1145/2072298.2071969

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Liu, P., Feng, M., Zheng, F., Zhang, J., Luo, P., Luo, J., & Xu, C. (2025). Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 1. https://doi.org/10.1109/TCSVT.2025.3566695

Tools, U. (2025). Design tool survey.

UserTesting (2021). *UserTesting: The human insight platform*. https://www.usertesting.com/.

Van Berkel, N., Opie, J., Ahmad, O. F., Lovat, L., Stoyanov, D., & Blandford, A. (2022). Initial responses to false positives in ai-supported continuous interactions: A colonoscopy case study. *ACM Transactions on Interactive Intelligent Systems*, *12*(1), 1–18.). https://doi.org/10.1145/3480247

Wang, I., Narayana, P., Smith, J., Draper, B., Beveridge, R., & Ruiz, J. (2018). *Easel: Easy automatic segmentation event labeler* [Paper presentation]. Proc. ACM IUI, IUI '18, New York, NY, USA (pp. 595–599). Association for Computing Machinery.

Wang, J., Qiu, K., Peng, H., Fu, J., & Zhu, J. (2019). *Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance* [Paper presentation]. Proc. ACM MM, New York, NY, USA (pp. 374–382). Association for Computing Machinery. https://doi.org/10.1145/3343031.3350910

Wang, O., Schroers, C., Zimmer, H., Gross, M., & Sorkine-Hornung, A. (2014). Videosnapping: Interactive synchronization of multiple videos. *ACM Transactions on Graphics*, *33*(4), 1–10.). https://doi.org/10.1145/2601097.2601208

Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., & Bansal, M. (2025). Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 3272–3283).

Xia, J., Singh, V., Wilson, D., & Latulipe, C. (2014). *Exploring the design space of multiple video interaction* [Paper presentation]. Proc. ACM NordiCHI, NordiCHI '14, New York, NY, USA (pp. 276–285) Association for Computing Machinery. https://doi.org/10.1145/2639189.2639247

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). *Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding* [Paper presentation]. Proc. ACM IUI, IUI '23 Companion, New York, NY, USA (pp. 75–78). Association for Computing Machinery. https://doi.org/10.1145/3581754.3584136

Zhu, G., Huang, Q., Xu, C., Rui, Y., Jiang, S., Gao, W., & Yao, H. (2007). *Trajectory based event tactics analysis in broadcast sports video* [Paper presentation]. Proc. ACM Int. Conf. Multimed., MM '07, New York, NY, USA (pp. 58–67) Association for Computing Machinery. https://doi.org/10.1145/1291233.1291250

About the authors

Luyao Shen received MA in Artistic Design from the University of Science and Technology Beijing in 2020. She is conducting her PhD in Computational Media and Arts (CMA) at HKUST(GZ). Her research interests include human-computer interaction and user experience design.

Qing Shi is an MPhil student at CMA Thrust of HKUST(GZ). His recent research interests include visualization and visual analytics, and human-computer interaction.

Emily Kuang received BASc in Biomedical Engineering from the University of Waterloo in 2020. She completed a PhD in Computing and Information Sciences from the Rochester Institute of Technology in 2025, and is currently an Assistant Professor at York University. Her research interests include human-AI collaboration and accessibility.

Linjie Qiu received BEng degree in digital media technology from Xiamen University in 2024. He is currently a MPhil student at the APEX lab in the thrust of CMA at HKUST(GZ). Her research interests include human-computer interaction, virtual reality and augmented reality.

Shixu Zhou received Bsc from the Hong Kong University of Science and Technology in 2022. He also received the MPhil degree from HKUST(GZ). He is currently a PhD student in the Computer Science and Engineering Department at HKUST. His research interests include human-computer interaction, human-AI collaboration, and AI for education.

Pan Hui received PhD from the University of Cambridge. He is a Chair Professor and Director of the Center for Metaverse and Computational Creativity at the Hong Kong University of Science and Technology (Guangzhou). Additionally, he holds the Nokia Chair in Data Science at the University of Helsinki.

Mingming Fan is an Associate Professor at Hong Kong University of Science and Technology (Guangzhou) and Hong Kong University of Science and Technology. He directs the Accessible & Pervasive User Experience (APEX) group to conduct research in the fields of Human-Computer Interaction, Aging and Accessibility, and VR/AR/XR.

Appendix A

Visual analytics tool incoporates ML-driven features to assist in identifying potential usability problems (Batch et al., 2024; Fan et al., 2022; Soure et al., 2022). This design offers several advantages:

- It is flexible to include additional interactive elements, such as a "filter" function to highlight desired data.
- It organizes information effectively by distributing data across various panels. In contrast, conversational AI assistant displays all information in a single chat thread, including potential usability problems, user queries, and AI responses).

However, the visual analytics tool also encounters notable drawbacks:

- The interface with timeline visualization appears information-rich, potentially leading to cognitive overload.
- The ML-driven data typically functions as a non-interactive display, lacking human-AI interaction.

Conversational AI assistant has been leveraged in prior studies to both proactively suggest potential usability problems and passively answer user inquiries (Kuang et al., 2023). This design has several advantages:

- Conversation is emerging as a key mode of human-computer interaction.
- Provide timely responses to user questions and elaborate on AI suggestions.

Despite these benefits, the design also presents drawbacks:

- The conversational AI assistant has a fixed modality that lacks adaptability.
- Manual effort required. The use of AI algorithms to accurately detect usability problems and generate natural language responses remains challenging. Consequently, prior studies have employed the Wizard of Oz technique, which involves manually extracting information from usability testing videos in advance to implement this design (Kuang et al., 2024).

Appendix B

We recruited 5 users (2 females, 3 males) (Nielsen, 2023) to complete the following tasks on a mobile weather application: 1. Customize the home screen so that UV is visible at the top. 2. Add Copenhagen to the list of saved locations and change the nickname of this city to Vacation. 3. Find the weather forecast for Copenhagen on Nov 16. 4. Remove the nickname Vacation and reset it to the default name.

Then, we recruited another 5 users (3 females, 2 males) to complete the following tasks in a VR game application: 1. Select the Blocking game from the menu and play it. 2. Select the Volleyball game from the menu and play it. 3. Select the Skeet game from the menu and play it.

We tested *MultiUX* on these two additional datasets, and the results are shown in Table B1.

Table B1. *MultiUX*'s Performance on additional product types.

Product type	Precision	Recall	Alignment rate of category	Alignment rate of subcategory
Mobile App	81.8%	67.6%	83.9%	80.6%
VR App	85.8%	61%	81.3%	75%

Appendix C

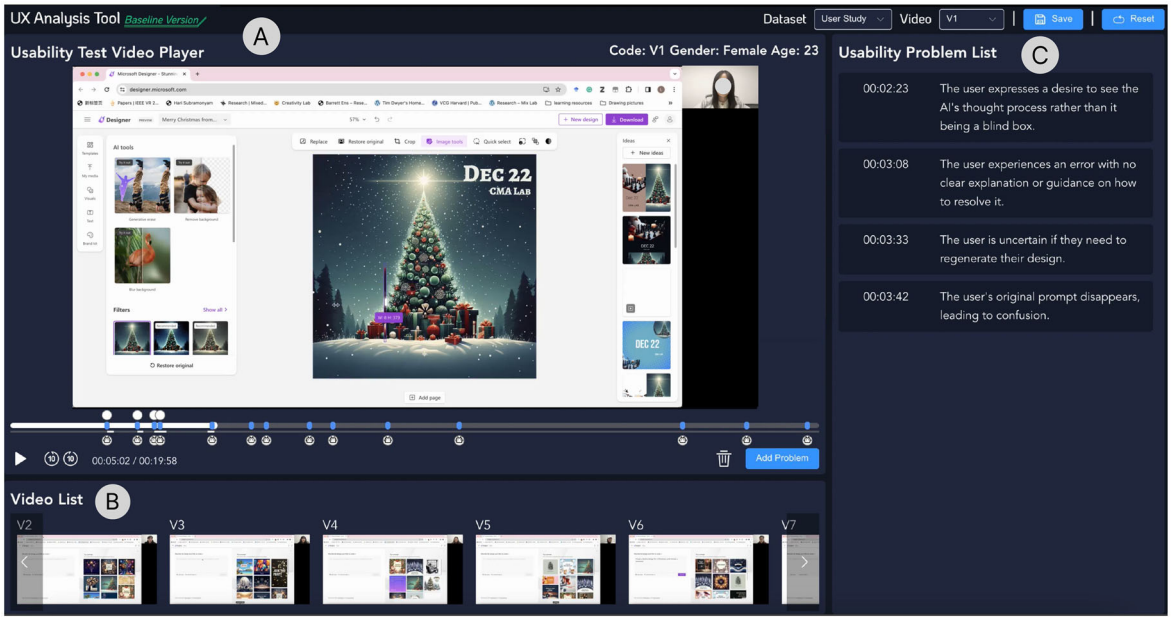

Figure C1. The baseline version includes (A) *usability test video player* for viewing videos and identifying usability problems, (B) *Video list* organized by participant number, (C) *Usability problem list* displaying all identified problems with their timestamps.