Multi-Modal User Modeling for Task Guidance: A Dataset for Real-Time Assistance with Stress and Interruption Dynamics

Amal Hashky* University of Florida
Benjamin Rheault† University of Florida
Ahmed Rageeb Ahsan‡ University of Florida
Brett Benda§ University of Florida
Tyler Audino ¶ University of Florida
Samuel Lonneman ‖ University of Florida
Eric D. Ragan** University of Florida

Abstract

Integrating artificial intelligence (AI) with wearable technology and the power of augmented reality (AR) holds great potential for enabling real-time task assistance. Through wearable sensors and cameras, it is possible to monitor both the state of the physical world and the biometric status of a human operator, which can allow the AI to offer adaptive assistance tailored to match the user's needs and optimize aid. Developing AI models for such purposes requires training data that not only describes users' cognitive state during tasks but also accounts for their responses to the interactive system during the task. We introduce the Multi-Modal User Modeling for Task Guidance (MUMTG) to support the development of AI models for such purposes. The dataset is created through human-subjects studies with users performing search and assembly procedures with help from virtual instructions provided by either a head-worn AR headset or a monitor screen. The data-collection study uses a game-like scenario to guide participants through six guided tasks that vary in difficulty.

Within each group, we manipulated task duration and induced several stress triggers to increase the task cognitive demand for three tasks. The dataset includes physiological data such as electrodermal activity, temperature, heart rate, pupil dilation, and gaze. We also collected subjective self-report ratings regarding task workload and emotional responses after each task. We offer this rich dataset as a valuable resource to facilitate the development of user models for task guidance in highly demanding contexts.

Index Terms:
Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; Applied computing—Education—Computer-assisted instruction

1 Introduction

Personalization of digital assistants and real-time information displays can dynamically adapt system output to align with users' needs, but developing such technology requires a deep understanding of their behaviors and interactions within their environment to create a model representation of the user [5]. Further, integrating artificial intelligence (AI) with wearable technology and the power of Augmented Reality (AR) holds great potential for enabling personalized real-time task assistance [21]. Through wearable sensors and cameras, it is possible to monitor both the state of the physical world and the biometric status of a human operator, which can allow the AI to offer adaptive assistance tailored to match the user's needs and optimize aid [18].

Real-time monitoring of user stress and cognitive state could help identify moments where users need more information in times of uncertainty, less information during periods of overload, or a different type of information to improve user understanding. Adapting to stress is relevant for many common cases that require cognitively demanding tasks under time constraints and high workloads—including multitasking scenarios in daily life as well as professional contexts. For instance, on the more extreme ends, medics often find themselves in high-stakes scenes where immediate response is crucial, and the margin for error is slim. Similarly, military personnel across different fields operate in contexts that may require them to perform sophisticated procedures over extended periods. Such scenarios characterized by intense cognitive demand can negatively impact their performance [1, 6, 16], often manifesting in emotional responses such as frustration and stress [3, 24]. Real-time task guidance can also be beneficial in less critical but still impactful scenarios. For example, chefs in busy kitchens could leverage such guidance to enhance coordination and execution efficiency. In any of these scenarios, where precision and timing are crucial, the ability to monitor and adapt to an individual's cognitive state can significantly impact outcomes, ensuring high-quality performance and reducing the potential for error.

As a basis for implementing personalization techniques and algorithms, sample data is needed from user performance and various sensors from different use cases; human-computer researchers call this user modeling [5]. Furthermore, machine learning researchers can leverage data-driven statistical techniques to learn mathematical models [19, 28] that can classify users based on specific categories or predict their psychological status [12] to provide adaptive model-based personalization. Developing AI models for such purposes requires training data that not only describes users' cognitive state during tasks but also accounts for their responses to the interactive system during the task.

Our research provides a new multidimensional dataset of human behaviors, cognitive state, and biometric signals as they work through a task procedure with the aid of digitally-presented instructions under pressure. The Multi-Modal User Modeling for Task Guidance (MUMTG) dataset includes a range of physiological responses and subjective ratings as time-series data. In a game-like scenario, participants engaged in guided tasks that vary in difficulty under stressful and calm conditions. Participants were tasked with a interpreting instructions, object search, and construction with building blocks as the scenario progressed. The stressful tasks included various stress-inducing elements (stress triggers), including incomplete instructions, time limitations, and interruptions [15] provoking participants to complete the activity under pressure. On the other hand, the relaxed (unstressed) conditions did not include any stressors, allowing participants to engage in the tasks without any time

*e-mail: ahashky@ufl.edu
†e-mail: brheault@ufl.edu
‡e-mail: ahmedrageebahsan@ufl.edu
§e-mail: brett.benda@ufl.edu
¶e-mail: tyler.audino@ufl.edu
‖e-mail: slonneman@ufl.edu
**e-mail: eragan@ufl.edu

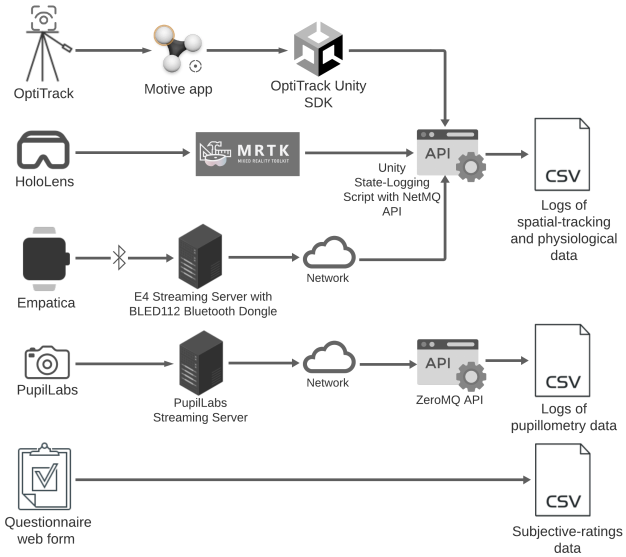

Figure 1: The diagram illustrates the integration of various technologies and the processing workflow for our multi-modal data collection.

Figure 2: Part of the physical game board for the user task

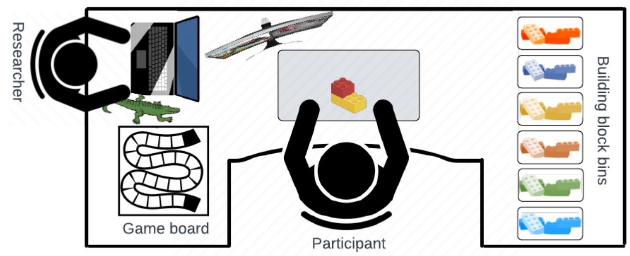

Figure 3: The physical study setup

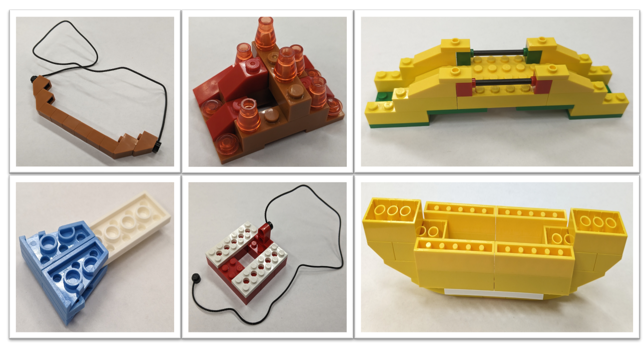

Figure 4: The builds ordered by level of difficulty as low, medium, and high from left to right. The top three are from the unstressed conditions, whereas the bottom three are from the stressed conditions.

restrictions or unanticipated challenges. The data collection studies account for two modes of real-time information display: headworn AR and on a monitor screen.

2 Data Collection

2.1 System

Our study employed a multifaceted approach to data collection, including spatial tracking, biometric data, and subjective feedback data. The diagram in Figure 1 illustrates the multi-modal data collection and processing workflow for the our study. It shows the integration of various technologies: the OptiTrack tracking system and Microsoft HoloLens headset for spatial tracking, the Empatica wristband for physiological data, and the PupilLabs camera for eye-tracking data. Data from these devices are streamed to servers and processed via different APIs (OptiTrack Unity SDK, NetMQ, and ZeroMQ), resulting in the compilation of logs in CSV format for spatial tracking, physiological, and pupillometry data separately. The subjective ratings are collected through web-based forms.

2.2 User Task

In a game-like scenario, participants engaged in guided tasks that vary in difficulty. Each participant had to complete six tasks with building blocks; two were simple with a few steps, two were more challenging with several steps ranging between 4–7, and the other two were complex builds with more than seven steps. Figure 2 shows part of the physical game board used in the study, and Figure 3 shows the physical setup. Figure 4 shows the constructed objects the participants had to complete during the six tasks, the images ordered by level of difficulty from left to right as low, medium, and high. The top three were from the unstressed conditions, whereas the bottom three were from the stressed conditions. Building-block assembly tasks are inherently modular and scalable, making them ideal for designing tasks varying in complexity. They require spatial reasoning, fine motor skills, and the ability to follow detailed instructions, making them suitable for simulating real-world tasks that require individuals to manage and adapt to varying workload activities. Further, the sequential nature of block assembly lends itself to the study of task assistance with the need for step-by-step instructions that can be presented through various formats. Also important for data collection, block building is an easily-understandable task for general study participants without the need for specialized expertise.

We designed the study to collect data under stressful and unstressed conditions by splitting the game into two halves, each consisting of three building tasks. During the first half, there was no chance of failure and no complications. For the second half, each task had a time limit, additional complications (stress triggers), and consequences for failure.

2.2.1 Stress Triggers

Each task in the second half of the game had one or two complications meant to provoke participants to complete the activity under pressure. Participants had to both complete the build and deal with the complications within the time limit. Table 1 shows a summary of the tasks with their associated level of difficulty and stress triggers. These stressors are designed to increase task workload, such as cog-

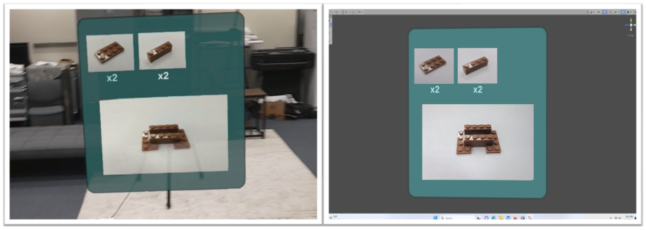

Figure 5: The instructions panel through the HoloLens headset in the AR condition (left) and on a monitor screen for the Desktop monitor condition (right).

nitive and temporal demands, simulating the multifaceted nature of challenges encountered in high-cognitive demand real-world tasks.

For task 4, the build instructions were given out-of-order, forcing participants to move forward and backward through the steps to figure out the correct sequence and guess in which order they should be done. This **out-of-order instruction** stressor simulates scenarios where information may not always be presented in a linear or logical sequence, adding extra strain on working memory to interpret the presented steps, solve spatial problems, and fix problems due to poor instructions.

As another stressor, participants were also given a **memorization task** at the end of task 4. For this stressor, they had to repeat five names given to them at the beginning of the study (approximately 30–45 minutes before reaching this task). This stressor was chosen to add an additional cognitive load, requiring individuals to retain and recall information, while also evoking a sense of frustration stemming from the need to remember names to avoid penalty in the game scenario.

For task 5, we introduced the stressor of **text form instructions** for the building blocks assembly. Here, the build instructions were given in text form instead of as images, making them difficult to interpret. This stressor was designed to increase cognitive load by requiring additional mental effort to convert textual information into spatial constructs, mirroring situations where instructions are unavailable in the most convenient format. Additionally, in the middle of the task, we interrupted participants and asked them to find an item on the game board. We added this stressor of **interruption** to reflect scenarios where individuals are often required to shift attention between tasks or attend to distractions from their environment while focusing on completing a specific task.

For task 6, the build instructions called for a piece purposely not included in the corresponding bin. Instead, participants had to search for the instructed item from among the other elements around them. The instructed game piece was hidden among the numerous colorful plastic decorative game setting, and participants had to visually scan the space and account for occlusion from other objects. The **hidden required piece** stressor was chosen to mimic real-life cases where it is difficult to keep track of specific items or tools, and cluttered workspaces introduce complexity for locating key items at inopportune times.

In addition to all the above stressors, we added **time pressure** in each of the final three tasks. Participants had to complete the tasks within a specific duration of time in order to avoid a penalty in the game progress, though the penalty did not directly affect the instructed building tasks. We added this stressor to create a sense of urgency and stress by putting pressure on individuals to manage their time efficiently while correctly working through the tasks.

2.3 Instruction Conditions

We used a between-subjects study design with two conditions. One group received instructions through a monitor screen, while the other received the instructions through the HoloLens AR display.

Table 1: Summary of the tasks with their associated level of difficulty and stress triggers.

Task	Difficulty Level	Stress Triggers
1	Medium	None
2	Low	None
3	High	None
4	High	• Out of order instructions • Memory task • Time pressure
5	Low	• Text form instructions • Interruption • Time pressure
6	Medium	• Hidden required piece • Time pressure

For consistent data collection, all participants wore the headset for the entirety of the study, but those in the desktop monitor condition did not have anything displayed through the AR system.

Participants in the AR condition also received cues to help them complete the challenges. For each step, an arrow was displayed through the AR display over the bin(s) that contained the needed piece(s). We contrasted the instructions delivery mode between augmented reality (AR) and traditional monitor screen display to assess the potential of immersive technology in facilitating following instructions with visual cues compared to the conventional format of static screens that would also require several physical shifts between the screen and the building space.

2.4 User Interface

We used Unity to create the user interface for the task. The interface consisted of an instruction panel containing directions for completing a single build step. Images of the required pieces for the step (a maximum of three) were displayed at the top, with a number below each image denoting how many of each piece is needed.

Below the required pieces is shown an image of the completed step, with all the pieces assembled. We designed each step to be a small change from the prior one to clarify the construction process.

Users in the AR condition viewed the instruction panel through the headset. For the desktop monitor condition the instructions were displayed on a monitor screen directly in front of the participant, an example is shown in Figure 5.

2.5 Procedure

The study was reviewed and approved by the university's ethics review board (Institutional Review Board; IRB) for protection and respect of participants. Participation in this study was voluntary. Upon participants' arrival, participants were comprehensively briefed about the study's objectives, procedures, and the sensor devices used for data collection. Participants were optionally offered extra credit in approved courses as compensation.

We conducted the study in a controlled lab setting, with individual sessions for each participant to maintain privacy and minimize any potential discomfort or stress. The dataset does not include any personal identifiers, thereby preserving the anonymity and ensuring the privacy of all participants. Following the initial briefing, participants were equipped with a headset, tracking gloves, and wristband and were asked to complete two preliminary activities. The first was a practice building task to allow them to familiarize themselves with the process of following the instructions to build an object. The second was a color detection task in which participants had to watch

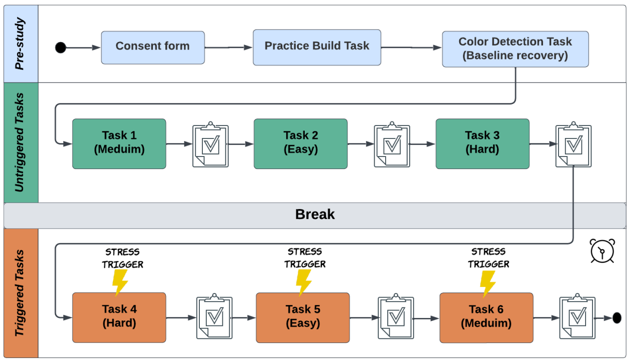

Figure 6: Procedure overview

randomly colored squares presented for 10 seconds and count the number of yellow squares passively. This task aimed to regulate their physiological responses to baseline levels [11] before beginning the non-stressed region of the game.

For the game, the participant was tasked with completing the three builds of the unstressed region before being offered a short break and then completing the three builds of the stressed region. After each task, the participant filled out the NASA Task Load TLX [8] and PANAS-X [25] questionnaires. Figure 6 show a visual summary of the study procedure.

3 Data Preprocessing and Labeling

The output of the data collection workflow, shown in Figure 1, included three CSV files: 1) spatial tracking and physiological data logs, 2) pupillometry data logs, and 3) subjective rating data. We crafted the dataset by following a structured preprocessing workflow. For each step, we grouped the records by task for each participant to guarantee the integrity of the results. Initially, we cleaned the spatial tracking and physiological data—this included extracting unique identifiers, synchronizing timestamps, filtering out irrelevant data falling out-of-tasks range duration, and encoding textual information. Additionally, we resampled the pupillometry data frequency to 60 Hz to match the spatial tracking and physiological data frequency. Following this, we integrated both logs based on timestamps, resulting in one log combining spatial tracking and biometric data. Finally, we calculated the NASA TLX and PANAS-X scores in the questionnaire responses, adhering to the standardized scoring guidelines of each instrument to derive multi-labels outcomes for the spatial tracking and physiological data.

4 Dataset

We elaborate on the different types of data collected during the study, and show descriptive statistics and figures from a portion of the data analyzed so far (20 participants). The sample described in this paper is a subset of data collected and processed to date, and we are continuing to add to the dataset as we collect and process new data. The (MUMTG) dataset presented in this paper is publicly available in Mendeley Data repository [9].

4.1 Spatial Tracking Data

Head and hand movement patterns can offer valuable insights into non-verbal cues and physical manifestations of cognitive load and emotional status.

We captured several tracking data for the participants' heads and hands while executing the tasks. While one group of participants received the instructions for the tasks on a monitor screen; they were also equipped with the HoloLens for consistent data collection across all participants. The captured tracking data includes positional and

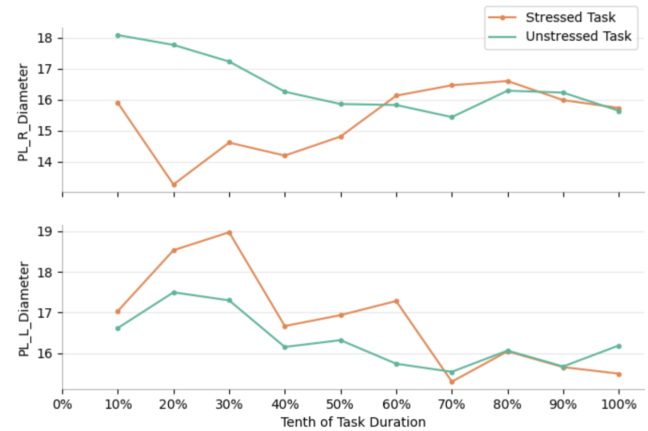

Figure 7: Patterns of pupil diameter dilation (in image pixels) in response to task engagement. Displayed are the trends of right and left pupil diameters throughout the task's progression, comparing a medium difficulty task between stressed and unstressed conditions.

Table 2: Comparative analysis of pupillometry metrics for right and left pupils during unstressed and stressed tasks.

		Unstressed Tasks		Stressed Tasks	
		Mean	SD	Mean	SD
Right pupil	Diameter	13.84	18.09	13.87	17.36
	Axes_X	10.49	12.65	10.14	12.86
	Axes_Y	13.84	18.09	13.87	17.36
	Angle	-13.7	83.02	-18.54	81.57
Left pupil	Diameter	14.71	16.82	14.76	25.68
	Axes_X	11.11	12.2	10.9	12.75
	Axes_Y	14.71	16.82	14.76	25.68
	Angle	-0.25	88.1	-4.88	88.43

rotational coordinates denoted as 'x,' 'y,' and 'z' values, providing a three-dimensional positional matrix for various reference points for the head and hands. The same movements were logged with an OptiTrack Prime 13W camera system with reflective markers observed by a set of infrared cameras to determine positions and orientations. This has accuracy benefits over the computer-vision based approach used by the Microsoft Hololens, and also can track the users hands while outside their view (the Hololens is limited to tracking hands only when its internal cameras can see them). We note that the coordinate systems do not align between the two systems (e.g., the origin of one space differs from the origin of the other).

4.2 Biometric Data

Research in psychology has shown that biometric data is a rich resource for deducing various physiological and psychological states, such as cognitive demands and emotional states (e.g. [17, 22, 27]). Based on the psychology and computer science literature, we selected specific biometric measures associated with cognitive demand and emotional states that we show below.

Pupillometry Data Studies have demonstrated a correlation between ocular metrics and various psychological states, including cognitive workload, frustration, and stress levels. [7, 10, 13]. We

Table 3: Comparative analysis of physiological parameters across unstressed and stressed tasks.

	Unstressed Tasks		Stressed Tasks	
Measure	Mean	SD	Mean	SD
Temperature (°C)	30.42	1.16	30.77	1.15
HR (BPM)	103.2	33.87	107.65	35.25
GSR (μS)	0.66	1.09	1.05	1.34
BVP (μV)	0.02	13.58	-0.01	14.22
IBI (ms)	0.66	0.25	0.62	0.22

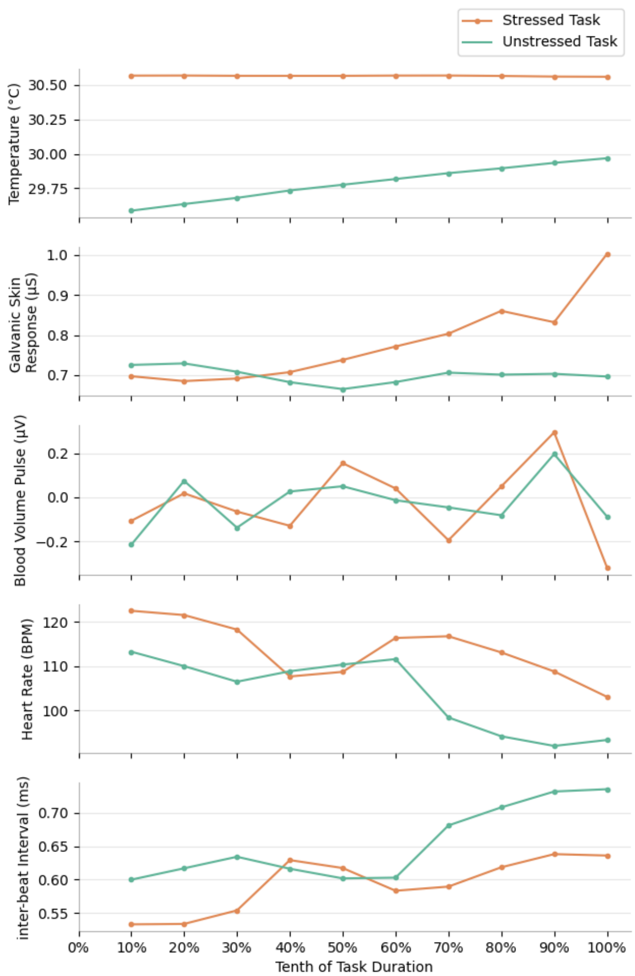

Figure 8: Multifaceted physiological responses over task duration. This graph displays changes in body temperature, galvanic skin response, blood volume pulse, heart rate, and inter-beat interval through different stages of task execution, comparing a medium difficulty task between stressed and unstressed conditions.

captured various eye measurements, focusing on the geometric and positional characteristics of the pupils. Including the normalized x-y coordinates locating the pupil within the eye camera's image frame and the pupils' diameter in image pixels. For detailed pupil morphology, we captured the ellipse's central x-y coordinates, the lengths of its major and minor axes, and the axial orientation (angle). Even in the absence of images or videos, the significance of such dynamic, continuous data lies in its ability to provide quantitative insights into human cognitive and physiological states; for example, changes in pupil diameter can indicate the cognitive load exerted by individuals during different tasks.

Figure 7 illustrates an example of the dynamic changes in pupil diameter (in image pixels) in response to task conditions and specific stress triggers, highlighting differences in ocular response across two tasks with medium difficulty in stressed vs. unstressed conditions; the task duration is segmented into deciles to facilitate comparison between the two tasks as they differ in duration. We also show the aggregated mean and standard deviation (SD) for pupil diameter (in pixel images), x and y axes coordinates of the major axis, and the rotation angle under stressed and unstressed conditions in Table 2.

Physiological Data Previous studies have shown that physiological data are linked to various emotional states, such as frustration and anxiety [2, 20, 23]. We gathered various metrics measuring participants' physiological responses while executing the tasks. These include body temperature, the galvanic skin response, blood volume pulse, heart rate, and inter-beat interval. This raw data provides a significant resource for extracting several insightful features to indicate changes in users' nervous system response, such as temperature variability, galvanic skin response (GSR) frequency, pulse rate variability, heart rate variability, and IBI variability. Table 3 presents the mean values and standard deviations (SD) for the above data, comparing the measurements obtained during tasks with stressors to those without.

Figure 8 shows an example of the progression of physiological metrics across two tasks with medium difficulty (stressed vs. unstressed); the task duration is segmented into deciles to facilitate comparison between the two tasks as they differ in duration.

4.3 Subjective Ratings

Existing research shows that self-reported measurements correlate with cognitive demand and emotions [4, 26], Therefore, we asked participants to provide their subjective ratings about the task workload experience and their emotions after each task. We utilized two well-known, established instruments: the NASA Task Load Index (TLX) [8] and the Positive and Negative Affect Schedule PANAS-X [25]; after computing the scores, we derived several outcomes that we used to produce a multi-label dataset. These are NASA Weighted Average Scores, Positive Affect Scores, and Negative Affect Scores.

We run several statistical tests on the ratings data to check for differences across the instruction delivery conditions (AR vs. desktop monitor) and the trial conditions (stressed vs. unstressed). The results were statistically insignificant across the AR and desktop monitor conditions. For the stressed and unstressed conditions, Figure 9 details the findings from the Wilcoxon Rank Test. It highlights the influence of stress triggers on both the NASA Task Load Index and the affect scores. The box-plots illustrate the distribution of NASA Weighted Average, Positive Affect, and Negative Affect scores grouped by the levels of task difficulty (low, medium, high). The findings from wilcoxon-signed rank test reveal significant differences ($p < 0.05$) between stressed and unstressed tasks across all these measures except for the positive affect scores in the low difficulty task with $p = 0.06$.

5 Conclusion

This paper describes the collection of the Multi-Modal User Modeling for Task Guidance (MUMTG) dataset, comprising raw biometric, spatial tracking data, and subjective ratings from various tasks performed by 20 participants. The scenario for data collection simulates a variety of cognitively demanding tasks combining: search and assembly tasks with physical objects; information processing through digitally presented instructions; and state monitoring of the physical game setup along with external stressors. We systematically manipulated task difficulty and various stress triggers such as interruptions,

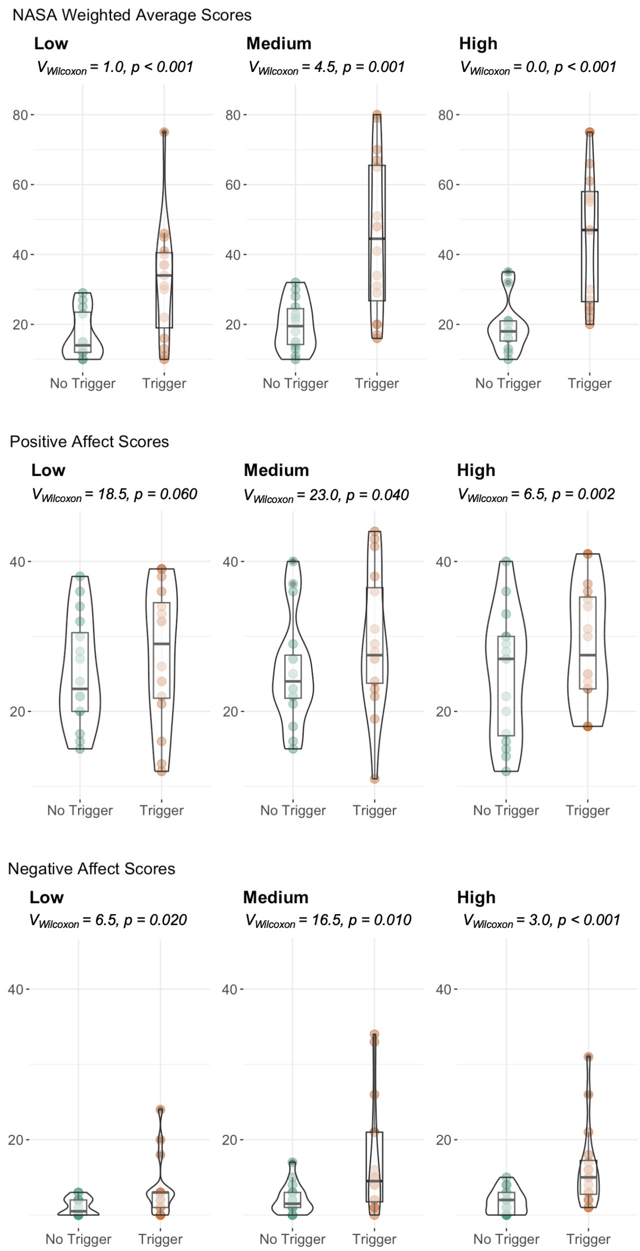

Figure 9: Impact of stress triggers on NASA Task Load Index and Affect scores. The boxplots represent the distribution of NASA Weighted Average Scores and Positive and Negative Affect Scores under varying task difficulties (Low, Medium, High). The figure compares the scores with and without stressors, indicating significant differences as per the Wilcoxon-signed rank test results.

time pressure, and cognitively demanding instructions. We also manipulated the instruction delivery to increase the variability of the data.

The current dataset is a work in progress, and we continue to run participants to increase the available sample for future researchers. With feature engineering, researchers can harness this data for machine learning and creating AI systems capable of providing real-time, adaptive task guidance. Our dataset is intended to supplement existing datasets in the domain. For example, the dataset referenced in [14] includes physiological metrics and subjective task ratings gathered from participants engaged in tasks related to knowledge work under conditions manipulated by stressors, which follows a motivation similar to our own for enabling user modeling. Importantly, our work's implementation across two display modalities (AR and monitor screens) is important and novel due to the ability to differentiate users' behavioral changes due to different types of display functionality that is also likely to affect behaviors. For example, when information is presented on a screen in the physical world, we can expect users to physically turn to access the information, while this would not be the case in AR. In AR, we might expect more severe changes in eye movements and pupil reactions for abrupt alterations in the visual head-worn overlay. Through future work with our dataset, research and development of AI models for adaptive information presentation will be able to account for reactions related to display differences more accurately. Through future analysis of the dataset, we also hope to inform the need for testing different display types and interaction modalities by improving the understanding of their effects on human responses, which is important for user modeling.

Acknowledgments

This work was supported in part by the DARPA Perceptually-enabled Task Guidance (PTG) Program under HR00112220005.

References

[1] N. C. Armstrong, S. J. R. Smith, D. Risius, D. Doyle, S. L. Wardle, J. P. Greeves, J. R. House, M. Tipton, and M. Lomax. Cognitive performance of military men and women during prolonged load carriage. *BMJ Mil Health*, 169(1):37–45, 2023. Publisher: British Medical Journal Publishing Group.

[2] B. B. Arnetz and R. Ekman, eds. *Stress in Health and Disease*. Wiley, 1 ed., July 2006. doi: 10.1002/3527609156

[3] I. L. Child and I. K. Waterhouse. Frustration and the quality of performance: II. A theoretical statement. *Psychological Review*, 60(2):127, 1953. Publisher: American Psychological Association.

[4] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, pp. 49–54. ACM, Los Angeles California, July 2010. doi: 10.1145/1836135.1836143

[5] G. Fischer. User Modeling in Human–Computer Interaction. *User Modeling and User-Adapted Interaction*, 11(1):65–86, Mar. 2001. doi: 10.1023/A:1011145532042

[6] B. Fu, W. Liu, and X. Liu. Influence of cognitive ability on task performance of dynamic decision making in military vehicles under different task complexity. *Cognition, Technology & Work*, 21(3):445–455, 2019. Publisher: Springer.

[7] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pp. 301–310. ACM, Copenhagen Denmark, Sept. 2010. doi: 10.1145/1864349.1864395

[8] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati, eds., *Advances in Psychology*, vol. 52 of *Human Mental Workload*, pp. 139–183. North-Holland, Jan. 1988. doi: 10.1016/S0166-4115(08)62386-9

[9] A. Hashky, B. Rheault, A. R. Ahsan, B. Benda, T. Audino, S. Lonneman, and E. Ragan. Multi-Modal User Modeling for Task Guidance (MUTMG): A Dataset for Real-Time Assistance with Stress and Interruption Dynamics. 1, Feb. 2024. Publisher: Mendeley Data. doi: 10.17632/b7c2h6cbc6.1

[10] S. T. Iqbal, X. S. Zheng, and B. P. Bailey. Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, pp. 1477–1480. ACM, Vienna Austria, Apr. 2004. doi: 10.1145/985921.986094

[11] J. R. Jennings, T. Kamarck, C. Stewart, M. Eddy, and P. Johnson. Alternate Cardiovascular Baseline Assessment Techniques: Vanilla or Resting Baseline. *Psychophysiology*, 29(6):742–750, 1992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1992.tb02052.x. doi: 10.1111/j.1469-8986.1992.tb02052.x

[12] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, Aug. 2007. doi: 10.1016/j.ijhcs.2007.02.003

[13] J. Klingner. Fixation-aligned pupillary response averaging. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*, p. 275. ACM Press, Austin, Texas, 2010. doi: 10.1145/1743666.1743732

[14] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij. The SWELL Knowledge Work Dataset for Stress and User Modeling Research. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 291–298. ACM, Istanbul Turkey, Nov. 2014. doi: 10.1145/2663204.2663257

[15] G. Mark, D. Gudith, and U. Klocke. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pp. 107–110. Association for Computing Machinery, New York, NY, USA, Apr. 2008. doi: 10.1145/1357054.1357072

[16] N. L. Miller, P. Matsangas, and L. G. Shattuck. Fatigue and its effect on performance in military environments. In *Performance under stress*, pp. 247–266. CRC Press, 2018.

[17] S. C. Müller and T. Fritz. Stuck and Frustrated or in Flow and Happy: Sensing Developers' Emotions and Progress. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, vol. 1, pp. 688–699, May 2015. ISSN: 1558-1225. doi: 10.1109/ICSE.2015.334

[18] M. Nardelli, S. Condino, S. Ghiasi, A. L. Callara, G. Rho, M. Carbone, V. Ferrari, E. P. Scilingo, and A. Greco. Recognizing AR-guided manual tasks through autonomic nervous system correlates: a preliminary study. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6, June 2020. doi: 10.1109/MeMeA49120.2020.9137309

[19] R. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, Oct. 2001. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. doi: 10.1109/34.954607

[20] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby. Anxiety detecting robotic system–towards implicit human-robot collaboration. *Robotica*, 22(1):85–95, 2004. Publisher: Cambridge University Press.

[21] P. Renner and T. Pfeiffer. Attention Guiding Using Augmented Reality in Complex Environments. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 771–772, Mar. 2018. doi: 10.1109/VR.2018.8446396

[22] P. Richter, T. Wagner, R. Heger, and G. Weise. Psychophysiological analysis of mental load during driving on rural roads-a quasi-experimental field study. *Ergonomics*, 41(5):593–609, May 1998. doi: 10.1080/001401398186775

[23] J. Scheirer, R. Fernandez, J. Klein, and R. W. Picard. Frustrating the user on purpose: a step toward building an affective computer. *Interacting with computers*, 14(2):93–118, 2002. Publisher: OUP.

[24] P. E. Spector. ORGANIZATIONAL FRUSTRATION: A MODEL AND REVIEW OF THE LITERATURE. *Personnel Psychology*, 31(4):815–829, Dec. 1978. doi: 10.1111/j.1744-6570.1978.tb02125.x

[25] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6):1063, 1988. Publisher: American Psychological Association.

[26] A. Weidemann and N. Russwinkel. Investigation of Frustration. In *Proceedings of Mensch und Computer 2019*, pp. 819–824. ACM, Hamburg Germany, Sept. 2019. doi: 10.1145/3340764.3344915

[27] G. F. Wilson. An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology*, 12(1):3–18, Jan. 2002. doi: 10.1207/S15327108IJAP1201_2

[28] I. Zukerman and D. W. Albrecht. Predictive Statistical Models for User Modeling. *User Modeling and User-Adapted Interaction*, 11(1):5–18, Mar. 2001. doi: 10.1023/A:1011175525451