A Universal Web-Based Tool for Multimodal Data Synchronization and Labeling

Nibraas Khan [1(✉)], Ruj Haan [1], Ingrid Shragge [2], Gabija Zilinskaite [2], Abigale Plunk [1], John Staubitz [2], Adithyan Rajaraman [2], Amy Weitlauf [2], and Nilanjan Sarkar [1]

[1] Vanderbilt University, Nashville TN 37235, USA
nibraas.a.khan@vanderbilt.edu
[2] Vanderbilt University Medical Center, Nashville TN 37232, USA

Abstract. Contemporary research methodologies demand sophisticated multimodal data integration, requiring precise synchronization of video recordings with diverse temporal data streams. However, researchers often struggle with the technical complexities of aligning and analyzing these data sources, leading to inefficient workflows and potential errors in temporal alignment. We present a universal web-based labeling tool designed to streamline the synchronization, visualization, and labeling of multimodal data across various research contexts. The tool provides an intuitive web-based interface that integrates multiple video streams, sensor data visualization, and flexible event coding capabilities while maintaining millisecond-level temporal precision through Unix timestamp synchronization. Key features include dynamic window management for multiple video streams, real-time audio mixing, interactive sensor data visualization, and a customizable labeling system supporting both point and duration observations. We demonstrate the tool's practical application through a behavioral research use case studying children with intellectual and developmental disabilities, highlighting its ability to manage complex multimodal data streams while maintaining temporal alignment and labeling accuracy. The tool's web-based architecture eliminates platform dependencies and installation requirements, while its open source nature ensures accessibility across research communities. By providing a free, standardized solution for multimodal data analysis, our tool addresses a critical need in modern research methodology, enabling researchers to focus on generating insights rather than managing technical complexities.

Keywords: Multimodal Data · Data Labeling · Data Synchronization · Research Methodology

1 Introduction

Modern research generates a wealth of data, integrating various data streams such as sensor readings, video recordings, and other modalities. Whether it is

M. Kurosu and A. Hashizume (Eds.): HCII 2025, LNCS 15771, pp. 254–264, 2025.
https://doi.org/10.1007/978-3-031-93965-5_18

evaluating manufacturing workflows, analyzing athletic performance, monitoring medical procedures, or studying classroom interactions, researchers face a common challenge: synchronizing and analyzing diverse modalities to uncover meaningful insights [14,25,27]. This process is inherently complex, requiring researchers to align and analyze disparate data sources, a task that is both time consuming and technically demanding [3,9].

A critical limitation in current workflows is the tools available for synchronized video and data analysis. These tools are often overly rigid, domain-specific, or require significant technical expertise to implement effectively [8,18]. Researchers frequently find themselves forced to piece together multiple solutions or develop custom tools to meet their needs [2,8,17]. For example, an engineer studying production line efficiency might use separate programs to view video recordings, examine sensor outputs, and log timestamps in spreadsheets-all while trying to align these datasets manually. This fragmented approach not only requires significant time but also increases the likelihood of errors in data alignment, compromising the quality of insights.

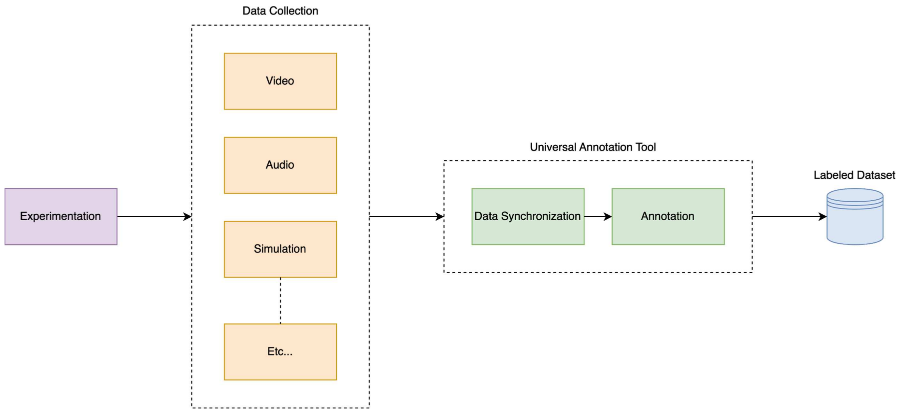


Fig. 1. Workflow abstraction from experimentation to data labeling using the Universal Coding Tool.

To address these challenges, we developed a universal web-based labeling tool designed to simplify the synchronization, visualization, and labeling of multimodal data. Our tool provides an intuitive interface that integrates video playback, temporal data visualization, and flexible event coding or annotating into a single platform. Researchers can work with multiple synchronized video streams, overlay relevant data, create custom coding schemes, and export standardized datasets ready for analysis. Figure 1 illustrates the workflow enabled by this tool.

The primary strength of our solution lies in its accessibility and adaptability. The tool runs entirely within a web browser, removing the need for complex installations and platform dependencies [24]. By focusing on user friendly design without sacrificing functionality, the tool enables researchers to concentrate on data analysis rather than navigating technical barriers.

In this paper, we describe the design and implementation of our universal labeling tool, demonstrating how it streamlines the analysis of synchronized video and data across a variety of research contexts. Through a behavioral research use case, we highlight the tool's practical application in managing multiple video streams, sensor data, and complex labeling workflows. By offering this free, open-source solution, we aim to address a significant need in modern research: an efficient, accurate, and adaptable approach to analyzing multimodal datasets. Our tool can be accessed here: https://rujh.github.io/data-visualization-app.

The remainder of this paper is structured as follows: Sect. 2 details the tool's architecture and key features for multimodal data synchronization and labeling. Section 3 provides a concrete use case in behavioral research, demonstrating its functionality in real-world scenarios. Finally, Sect. 4 discusses the tool's broader implications for research workflows and outlines directions for future development.

2 Design and Implementation

Our tool was developed using React Native, featuring a centralized synchronization system that ensures millisecond-level precision across diverse data components. By utilizing a unified time control mechanism, the architecture aligns multiple data streams with high accuracy, using Unix timestamps as the primary temporal reference. This section outlines the tool's design and technical capabilities. Each subsection highlights the strategies employed to enhance workflow efficiency and technical precision while providing researchers with an intuitive platform.

2.1 File Organization and Import Structure

A structured file organization system is essential for managing and synchronizing multimodal data accurately. Our tool requires designated directories for video, audio, and sensor data, ensuring consistency in data input. Although this setup introduces some constraints, it plays a crucial role in standardizing workflows. Sensitive data never leaves a researcher's system, as our tool operates entirely locally without relying on cloud storage or external servers, ensuring maximum privacy and data security. Figure 2 illustrates the recommended folder structure. Video files must include a timestamp in their filenames (e.g., 1234567890-1.mp4), which serves as the temporal reference point for synchronization with sensor data. Sensor files, on the other hand, can have arbitrary names but must include a Time column containing Unix timestamps to indicate when each sample was recorded. This approach eliminates the need for specifying sample rates manually, reduces ambiguity, and facilitates automatic alignment of video and sensor data without additional preprocessing.

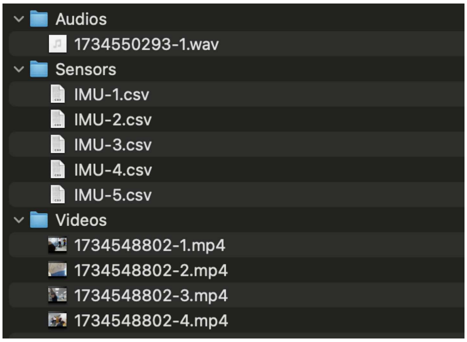


Fig. 2. Sample folder structure for data.

Video files must include a timestamp in their filenames (e.g., 1234567890-1.mp4), which serves as the temporal reference point for synchronization with sensor data. Sensor files, on the other hand, can have arbitrary names but must include a Time column containing Unix timestamps to indicate when each sample was recorded. This approach eliminates the need for specifying sample rates manually, reduces ambiguity, and facilitates automatic alignment of video and sensor data without additional preprocessing.

2.2 Video Management

Video analysis is critical in modern research methodologies, particularly in behavioral and social sciences [1, 11, 16, 20]. Our video management subsystem supports multiple synchronized video streams, enabling researchers to capture interactions from different perspectives. Complex scenarios often require simultaneous observation from various angles, and frame accurate synchronization ensures that all streams remain perfectly aligned even with high resolution recordings. Dynamic window management provides researchers with flexibility in their analytical workflows. Each video stream can be expanded into pop-out windows for detailed inspection, and these windows can be freely resized and repositioned to create a personalized workspace. Despite these movements, the system maintains absolute temporal synchronization. Navigation controls offer granular playback management. Play and pause actions apply uniformly across all streams, preserving contextual integrity. Researchers can navigate recordings with precision using 10-second forward or backward jumps to quickly locate moments of interest through direct timestamp entry (in HH:MM:SS format) and a video scrubber.

Individual volume and mute controls add another layer of flexibility. Researchers can isolate specific video streams while maintaining visual synchronization across all video feeds. For example, in a study on group dynamics, a researcher could focus on the video of a single participant while retaining the broader visual context.

2.3 Audio Integration

Audio data provides a nuanced and often underutilized dimension in research, offering critical contextual information [5,13,23,26]. Our audio integration subsystem supports multiple WAV audio sources, recognizing that many research scenarios require simultaneous audio captures from different perspectives. Real-time audio mixing provides granular control over sound sources. Researchers can dynamically adjust individual stream volumes, isolating specific audio channels while maintaining the broader acoustic context. This functionality is particularly useful in complex auditory environments, where multiple sound sources require focused analysis. Automatic start time detection eliminates the need for manual alignment by identifying the temporal origins of each audio stream.

2.4 Sensor Data Visualization

Sensor data visualization is a crucial component of research [6,7,10,21]. Our visualization subsystem provides standard approaches, including line graphs for time series data, scatter plots for variable relationships, and heatmaps for multi-dimensional datasets. The system is designed to support both standard and specialized visualizations, enabling the integration of advanced techniques such as 3D biomechanical reconstructions or neurological signal maps. This extensibility ensures the platform adapts to diverse research contexts and methodologies. Synchronization is a core feature, with all visualizations aligned via timestamp based mechanisms to maintain temporal consistency with video and audio streams. Additionally, interactive exploration enhances the analytical utility of the system. Researchers can manipulate visualizations to focus on specific intervals or data points, extracting detailed contextual information through overlays and tooltips.

2.5 Labeling System

Labels are the link between data and scientific insights [4,15,19,22]. Our observation system provides a flexible framework for capturing nuanced behaviors and events in complex research scenarios. It supports two key label types: event recording, which mark instantaneous events with millisecond precision, and duration observations, which track events unfolding over time. The system prioritizes customization and interactivity, allowing researchers to create tailored labeling schemas with custom names, metadata, and hierarchical categorizations. To enhance usability, duration buttons change color when activated, providing a clear visual reminder that an observation is still in progress. Observations can be added in real-time during playback or retrospectively, and the system enables seamless editing, refinement, and exploration of labels. Features like "go-to-time" navigation and active tracking of ongoing observations streamline the analysis of complex, overlapping events while allowing for quick edits, such as correcting erroneous behavior recordings. Additionally, the system displays timestamps and labels alongside the video, eliminating the need to scroll and ensuring precise alignment between observations and recorded events.

2.6 Session Management

Our session management system includes features like autosave, which continuously captures session states to prevent data loss from unexpected interruptions. The system supports comprehensive session exports, enabling researchers to archive their work as compressed packages that include video recordings, audio streams, sensor data, labels, and synchronization metadata. These exports maintain precise temporal relationships across all data sources, preserving the structural integrity of the research session. Multiple export formats, including CSV, offer compatibility with external tools, while optional encryption and error-checking mechanisms enhance data security and integrity.

3 Use Case

To demonstrate the practical applications of our universal labeling tool, we present a case study that builds upon recent advancements in behavioral research for children with intellectual and developmental disabilities (IDD). Research like Khan et al.'s pilot study on Real-time Early Agitation Capture Technology (REACT) highlights the complexity of collecting and analyzing behavioral data in autism research [12]. Applied Behavior Analysis requires meticulous Direct Behavior Observation (DBO), data collection, and analysis to monitor, assess, and respond to the behaviors of children who present with dangerous or interfering behavior. Therapists and researchers must precisely record specific behavioral events, engagement with activities or other individuals, and even potential distress signals. While technologies like REACT focus on data capture, our labeling tool addresses the need for comprehensive, precise DBO data labeling.

When analyzing the behavior of children with autism or IDD precise measurement tools are needed to record detailed, moment-to-moment events. Understanding the timing, sequence, frequency, and duration of behavior can all be necessary parameters for analysis, depending on the individual needs of the situation. Capturing brief instances of discrete behaviors such as throwing a toys, tracking duration of an ongoing behavior such as an extended social interaction with a therapist, or identifying environmental events that are tending to occur as reliable antecedents to a behavior of interest requires a sophisticated labeling approach that can synchronize multiple data sources, as well as valuable tools for inputting DBO data.

3.1 Data Loading and Options

Researchers begin their workflow by uploading a complete research session folder to the web-based platform. Upon successful upload, the tool automatically parses and organizes all data sources, ensuring consistency and accuracy across modalities. This includes video recordings, audio, sensor data, and other relevant inputs. An initial summary of the data is provided, allowing researchers to verify successful integration.

(a) Data loading page (b) Options page

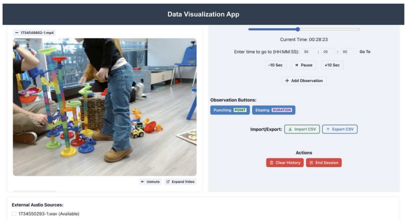
Fig. 3. Flow for loading data into the tool along with configuring sensor and labeling options.

The observation configuration process is highly adaptable, allowing researchers to define and record observable behaviors and environmental events relevant to their specific context, such as a therapy session, classroom routine, or community based observation. Researchers can mark the precise timing of the onset of a discrete behavior using an event recording approach to DBO, appropriate for instances of aggressive behavior that might start and end in 1–3 seconds, as well as duration based recording, that would highlight the amount of time that the child was eloping from an instructional area or a sequence of vocalizations This flexibility ensures that observation schemas align with a wide range of research objectives or therapeutic goals. The overall flow of this step is illustrated in Fig. 3.

3.2 Labeling Process

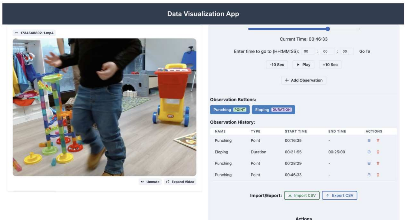
The labeling interface is designed to transform raw multimodal data into actionable behavioral insights. By providing a comprehensive view of the observation session, it enables researchers or clinicians to analyze interactions from multiple perspectives. Video streams from different angles capture detailed views of the therapy environment, allowing researchers to observe participant behavior and interactions simultaneously. Additionally, when the person intended to be the focal point of a video is not stationary and a camera cannot follow them, the possibility for integrating video from multiple angles serves to prevent gaps in an observation. Meanwhile, audio recordings from strategically placed microphones add crucial context, capturing subtle or low-volume verbal exchanges and environmental sounds that may not be fully visible in video footage or audible from a single microphone placement location.

Synchronized sensor data are shown with the visual and audio streams. Researchers or clinicians can fine-tune graph parameters, zoom into specific time windows, and compare sensor streams to uncover nuanced relationships. When the researcher initiates playback, the tool synchronizes all data sources, offering a detailed, moment-by-moment view of the participant's behavioral responses. The user interface is provided in Fig. 4.



(a) Videos with play control

(b) Raw data synced with all other modalities



(c) Sample labels

Fig. 4. Data labeling with synced modalities

3.3 Label Export

The data export functionality is designed to provide researchers and clinicians with flexible and efficient options for preserving and analyzing their behavioral coding data. Sessions can be exported in multiple formats, with CSV as the primary recommended format due to its versatility and compatibility with most analytical workflows. The exported CSV file contains a structured representation of the session's observations, providing a clear and customizable dataset. Table 1 illustrates the base column structure, which can be tailored to include or exclude specific data points based on research needs.

Table 1. Observation Data Table

Observation Name	Type	Timestamp	Start Time	End Time	Duration
Punching	Point	02:41:25.203Z	0:16:35		
Eloping	Duration	02:41:27.840Z	0:21:55	0:25:00	0:03:05
Punching	Point	02:41:22.486Z	0:28:29		
Punching	Point	02:41:34.012Z	0:46:33		

Beyond data preservation, the export system supports advanced workflows by allowing previously coded sessions to be re-imported into the tool. Researchers can reload their prior observations, resuming analysis where they left off. This feature is particularly valuable for iterative studies, enabling researchers to review and refine previous coding or conduct longitudinal analyses across multiple sessions. Additionally, this flexibility can expedite the process of correcting data collection errors, supporting data collector training, and checking for reliability across data collectors.

4 Conclusion

Our universal web-based labeling tool addresses significant challenges in multimodal research data analysis by offering an integrated, flexible, and intuitive platform. By combining video, audio, sensor data, and robust labeling capabilities within a single interface, the tool enables researchers and clinicians to work more efficiently and effectively.

A core strength of the tool lies in its synchronization, which ensures temporal consistency across diverse data sources. This feature eliminates the technical complexity often associated with aligning multimodal datasets. Additionally, the platform's visualization framework provides versatile options for representing sensor and behavioral data, supporting both standard and advanced visualization techniques. The labeling system offers a highly customizable framework for coding events. Researchers can create tailored labeling schemas, adjust observation categories, and manage detailed metadata, ensuring that the system aligns with their specific investigative needs. Beyond its analytical capabilities, the tool simplifies data preservation and sharing through export and import functionalities. Researchers can archive entire sessions, including labels and synchronization metadata, in compressed formats that maintain temporal integrity. These features support collaborative workflows, longitudinal studies, and data reanalysis, making the tool an invaluable resource for research teams and individual investigators alike.

Future development will focus on expanding the tool's capabilities to meet the evolving needs of the research community. Planned enhancements include advanced web services for sequential data analysis, more sophisticated visualization techniques, and the integration of machine learning-based analytical tools. As research methodologies become increasingly complex and data-intensive, tools that simplify multimodal analysis play an important role in supporting scientific progress. By providing accessible and advanced labeling and visualization technologies, our tool enhances the efficiency and accuracy of research workflows while fostering collaboration across disciplines. The impact of this tool extends beyond its technical functionality. It contributes to more streamlined and practical research practices, helping researchers manage the complexities of multimodal data analysis. Moreover, this tool and those that follow surmount one of the significant logistical obstacles that had previously made it impractical to integrate these diverse data sources that can bring greater impact and value when applied in this fashion. By integrating synchronization, visualization, and labeling processes, the tool offers a reliable solution for addressing challenging research questions, ultimately advancing knowledge across diverse scientific fields.

Acknowledgment. The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Foundation (grant number 2124002).

References

1. Ahmedt-Aristizabal, D., et al.: Deep learning approaches for seizure video analysis: a review. Epilepsy Behav. **154**, 109735 (2024)
2. Aljabri, M., AlAmir, M., AlGhamdi, M., Abdel-Mottaleb, M., Collado-Mesa, F.: Towards a better understanding of annotation tools for medical imaging: a survey. Multimedia Tools Appl., 1–35 (2022). https://doi.org/10.1007/s11042-022-12100-1
3. Barua, A., Ahmed, M.U., Begum, S.: A systematic literature review on multimodal machine learning: applications, challenges, gaps and future directions. IEEE Access **11**, 14804–14831 (2023)
4. Braun, D.: I beg to differ: how disagreement is handled in the annotation of legal machine learning datasets. Artif. Intell. Law **32**(3), 839–862 (2024)
5. Busquet, F., Efthymiou, F., Hildebrand, C.: Voice analytics in the wild: validity and predictive accuracy of common audio-recording devices. Behav. Res. Methods **56**(3), 2114–2134 (2024)
6. Cai, G., Zhang, Q., Liu, B., Jin, Z., Qian, J.: Deep learning-based recognition and visualization of human motion behavior. Acad. J. Sci. Technol. **10**(1), 50–55 (2024)
7. Fister Jr, I., Fister, I., Podgorelec, V., Salcedo Sanz, S., Holzinger, A.: NarmViz: a novel method for visualization of time series numerical association rules for smart agriculture. Expert Syst. **41**(3), e13503 (2024)
8. Friard, O., Gamba, M.: Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. Methods Ecol. Evol. **7**(11), 1325–1330 (2016)
9. Hakim, Z.I.A., Sarker, N.H., Singh, R.P., Paul, B., Dabouei, A., Xu, M.: Leveraging generative language models for weakly supervised sentence component analysis in video-language joint learning. arXiv preprint arXiv:2312.06699, 2023
10. Hashmi, S.J., Alabdullah, B., Al Mudawi, N., Algarni, A., Jalal, A., Liu, H.: Enhanced data mining and visualization of sensory-graph-modeled datasets through summarization. Sensors **24**(14), 4554 (2024)
11. Katona, Z., Mohammadi Ziabari, S.S., Karimi Nejadasl, F.: Marine: a computer vision model for detecting rare predator-prey interactions in animal videos. arXiv preprint arXiv:2407.18289 (2024)
12. Khan, N., et al.: Pilot study of a real-time early agitation capture technology (react) for children with intellectual and developmental disabilities. Digit. Health **10**, 20552076241287884 (2024)
13. Kumar, A., Singh, S.K., Bhardwaj, I. et al. Audio spectrogram analysis in IoT paradigm for the classification of psychological-emotional characteristics. Int. J. Inf. Tecnol., 1–11 (2024). https://doi.org/10.1007/s41870-024-02166-5
14. Liang, P.P., Zadeh, A., Morency, L.P.: Foundations & trends in multimodal machine learning: principles, challenges, and open questions. ACM Comput. Surv. **56**(10), 1–42 (2024)
15. López, J.A.H., Izquierdo, J.L.C., Cuadrado, J.S.: ModelSet: a labelled dataset of software models for machine learning. Sci. Comput. Programm. **231**, 103009 (2024)

16. Luxem, K., et al.: Open-source tools for behavioral video analysis: setup, methods, and best practices. Elife **12**, e79305 (2023)
17. Miao, R., Robert Toth, Yu., Zhou, A.M., Janowczyk, A.: Quick annotator: an open-source digital pathology based rapid image annotation tool. J. Pathol. Clin. Res. **7**(6), 542–547 (2021)
18. Mittal, M., Raheja, N.G.: Data Visualization and Storytelling with Tableau. CRC Press (2024)
19. Monarch, R.M.: Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI. Simon and Schuster (2021)
20. Nassauer, A., Legewie, N.M.: Video data analysis: a methodological frame for a novel research trend. Sociol. Methods Res. **50**(1), 135–174 (2021)
21. Qian, D., Zeng, H., Wenjie Cheng, Yu., Liu, T.B., Pan, J.: NeuroDM: decoding and visualizing human brain activity with EEG-guided diffusion model. Comput. Methods Programs Biomed. **251**, 108213 (2024)
22. Russo, S., et al.: The value of human data annotation for machine learning based anomaly detection in environmental systems. Water Res. **206**, 117695 (2021)
23. Sabry, A.H., Bashi, O.I.D., Ali, N.N., Al Kubaisi, Y.M.: Lung disease recognition methods using audio-based analysis with machine learning. Heliyon **10**, e26218 (2024)
24. Segun-Falade, O.D., Osundare, O.S., Kedi, W.E., Okeleke, P.A., Ijomah, T.I., Abdul-Azeez, O.Y.: Developing cross-platform software applications to enhance compatibility across devices and systems. Comput. Sci. IT Res. J. **5**(8) (2024)
25. Stöckl, H., Pflaeging, J.: Multimodal coherence revisited: notes on the move from theory to data in annotating print advertisements. Front. Commun. **7**, 900994 (2022)
26. Várkonyi, D.T., Bányai, D.T. and Varkonyi-Koczy, A.R.: Investigating traditional machine learning models and the utility of audio features for lightweight swarming prediction in beehives. Acta Polytech. Hung. **21**(10), 283–299 (2024)
27. Yang, X., et al.: Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. arXiv preprint arXiv:2305.10169 (2023)