Enriching Business Process Event Logs with Multimodal Evidence

Aleksandar Gavric(✉), Dominik Bork, and Henderik A. Proper

Business Informatics Group, TU Wien, Vienna, Austria
{aleksandar.gavric,dominik.bork,henderik.proper}@tuwien.ac.at

Abstract. Process mining uses data from event logs to understand which activities were undertaken, their timing, and the involved entities, providing a data trail for process analysis and improvement. However, a significant challenge involves ensuring that these logs accurately reflect the actual processes. Some processes leave few digital traces, and their event logs often lack details about manual and physical work that does not involve computers or simple sensors. We introduce the **B**usiness-knowledge **I**ntegration **Cycle**s (BICycle) method and *mm_proc_miner* tool to convert raw and unstructured data from various modalities, such as video, audio, and sensor data, into a structured and unified event log, while keeping human-in-the-loop. Our method analyzes the semantic distance between visible, audible, and textual evidence within a self-hosted joint embedding space. Our approach is designed to consider (1) preserving the privacy of evidence data, (2) achieving real-time performance and scalability, and (3) preventing AI hallucinations. We also publish a dataset consisting of over $2K$ processes with $16K$ steps to facilitate domain inference-related tasks. For the evaluation, we created a novel test dataset in the domain of DNA home kit testing, for which we can guarantee that it was not encountered during the training of the employed AI foundational models. We show positive insights in both event log enrichment with multimodal evidence and human-in-the-loop contribution.

Keywords: Event Log Creation · Event Log Completion · Event Log Quality Improvement · Artificial Intelligence · Multimodal data

1 Introduction

Consider an entertainment park, where the pre-designed pathways represent the official, mapped-out processes of an "idealized" visit to the park. These paths are laid out by planners with a specific flow in mind, directing the visitors (the participants of an entertainment process) on how they should navigate the space (or the entertainment business process). However, over time, the park's visitors may create a shortcut through the grass, a path not originally designed but formed out of convenience and efficiency. This real-life scenario serves as a perfect

Published by Springer Nature Switzerland AG 2025
E. Paja et al. (Eds.): PoEM 2024, LNBIP 538, pp. 175–191, 2025.
https://doi.org/10.1007/978-3-031-77908-4_11

metaphor for process discovery in the sense of process mining. This paper aims to explore the ways how processes are executed in practice and not blindly trusting the designed process, but rather by making the evidence data about process execution more complete.

Support for the design and improvement of business processes and realizing the benefits of information systems have been proposed by a broad spectrum of methods since the 1990s [18]. Pegoraro et al. [21] study process mining from uncertain event data and the result of automatically discovering process models and checking if event data conform to a certain model. Importantly, numerous authors [3] showed that tension between human involvement and task automation in work process management underscores the critical impact solutions to these identified problems will have on knowledge-intensive work processes with a conclusion that a better representation of the real-world context is crucial for process mining. To address the question of representing real-world contexts, this paper defines, implements, and evaluates the automatic creation of event logs for process mining from multiple modalities of data sources. In support of our motivation, the global video surveillance market, valued at USD 53.7 billion in 2023, is projected to reach USD 83.3 billion by 2028[1]. This growth is fueled by smart city initiatives and advancements in AI-driven video technology combined with the Internet of Things (IoT) across various sectors.

The research question addressed in this paper is: *How can multimodal evidence be effectively utilized to create or complete event logs?* This inquiry aims to explore the potential of combining different types of data sources, such as audio, video, text, and sensor data, to enhance the completeness and precision of event logs. The findings are expected to provide insights into improving event log generation and completion processes, ultimately contributing to more robust and dependable event documentation for the purposes of process mining. Consequently, this paper aims to contribute first techniques and a method toward realizing the vision of multimodal process mining [8].

The remainder of this paper is structured as follows. First, in Sect. 2, we introduce our general perspective on multimodal process evidence. This is followed, in Sect. 3, by an outline of related work. In Sect. 4, we then discuss the implementation of a tool for improving event log completeness with multimodal evidence, as well as the associated designed **B**usiness-knowledge **I**ntegration **Cycle**s (BICycle) method. Finally, before concluding, Sect. 5 reports on the evaluation of our solution. All the data of this research and the developed tool are available via: https://github.com/aleksandargavric/mm_proc_miner.

2 Perspective on Multimodal Process Evidence

Relying on a single modality (such as text) for event log creation is comparable to navigating a complex environment with only one sense. While valuable insights can be obtained, significant aspects of the process may remain obscured, leaving blind spots in our understanding and traceability of the process.

[1] https://www.marketsandmarkets.com/Market-Reports/video-surveillance-market-645.html.

Consider the scenario of assembling a piece of furniture; a process that involves various steps and interactions with multiple components. If we were to rely solely on written instructions (a single modality), we might miss out on the details of how different parts fit together, the angle of assembly, and the duration of each step that signals a correct assembly. These details might be better captured through video demonstrations (another modality), which provide a visual and auditory understanding of the assembly process. Similarly, sensor data from tools used in the assembly (yet another modality) could provide insights into the amount of force required for certain steps or the tactile feedback, offering a more complete picture of the process.

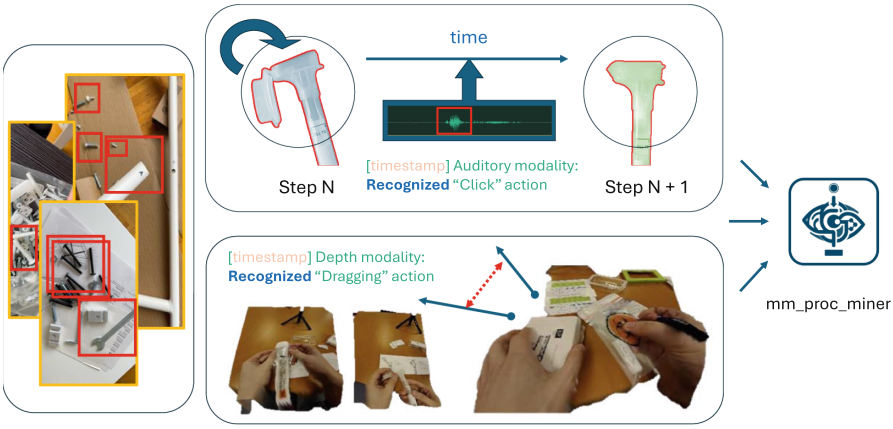

Fig. 1. An illustration of an effect from various modalities.

The integration of multiple modalities – textual, visual, auditory, and general sensor data – thus becomes crucial in uncovering the full spectrum of activities and interactions within a process. Each modality essentially offers a lens through which to view the process, revealing different aspects that may not be visible through other means. Just as a video can capture what written instructions cannot, sensor data can reveal details about the physical execution of a process that neither text nor video can capture.

In this paper, we show that by integrating diverse modalities, blind spots left by single-modal event log creation methods can be uncovered, ensuring a (1) privacy of evidence data, (2) performance and scalability, and (3) preventing AI hallucinations in the representation of processes in event logs. Our solution is illustrated in Fig. 1.

To define the research objective of this paper, we first formulate the research gap which we aim to contribute to.

Given a process P, let there be a set of N modalities $M = \{m_1, m_2, \ldots, m_N\}$, where each modality m_i provides a unique perspective of the process. Let, for $1 \leq i \leq N$, the set D_i represent the dataset of raw data (typically comprising unstructured data such as depth sensing maps, audio, and raw pixels of

video frames) associated with modality m_i, reporting on a sequence of events $E_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,O_i}\}$ observed through modality m_i. Each event $e_{i,j}$ in E_i is characterized by a tuple $(a_{i,j}, t_{i,j}, v_{i,j})$, where $a_{i,j}$ is the activity, $t_{i,j}$ is the timestamp, and $v_{i,j}$ is a vector of modality-specific attributes observed for the event.

Our objective is to construct a comprehensive event log *EL* that accurately represents the sequence and characteristics of activities in the process as observed across all modalities in M. Formally, we seek to optimize the following objective function:

$$\max_{EL} \sum_{i=1}^{N} \lambda_i \cdot \Phi(PM, D_i)$$

where $\Phi(EL, D_i)$ is a function that measures the fidelity of the event log *EL* with respect to the dataset D_i of modality m_i, and λ_i is a weighting factor that denotes the importance or reliability of modality m_i in the overall process understanding.

The challenge lies in accurately integrating the disparate and possibly conflicting information from different modalities to construct a process model that is both comprehensive and faithful to the observed data. This involves not only aligning events across modalities but also reconciling differences in the granularity, scale, and interpretation of the data.

To solve this optimization problem, we employ a multi-stage approach that first involves the alignment and fusion of multimodal datasets into a unified event log. Consequently, process discovery algorithms could be applied to this integrated dataset to construct the initial process model, which can be refined iteratively by evaluating its conformance against each modality-specific dataset and adjusting the model accordingly.

This paper proposes a novel method, termed **B**usiness-knowledge **I**ntegration **Cycle**s (BICycle), for addressing the challenges associated with digital traceability in manually-intensive business processes with limited digital footprints or those processes completely invisible to IT systems. The proposed approach aims to create process mining-ready event logs from videos and other unstructured data forms, with the goal of enhancing process monitoring and optimization. This research takes a domain-agnostic stance and involves human-in-the-loop design. It emphasizes the versatility and applicability of its findings across various fields. Integrating human-in-the-loop design ensures that human insights and feedback are integral to the iterative process, fostering a symbiotic relationship between technology and its users. This aims to not only enhance the relevance and usability of the solution but also align the solution with human (business-relevant) values and needs, incorporate diverse human perspectives, and make the solution user-centered.

To meet the above challenges, we have designed the BICycle method and developed a BICycle-enabled tool for the creation of event logs, which employs joint embedding space for different modalities, and shows a capability of identifying moments in video, audio, and unstructured sensor data representing activities for an event log.

3 Related Work

Process mining from multimodal data represents an emerging domain within process analytics, aiming at extracting meaningful process-related information from video data. This paper is positioned as a continuation of our earlier implementation described in [8], and described in [7].

We investigate work on improving the completeness of event logs is a critical area of research in process mining, as it ensures the reliability and accuracy of the insights derived from these logs. In particular, this involves addressing issues related to preventing "AI hallucinations" [14] that could lead to erroneous conclusions. The core challenge to improve the completeness of event logs is to enable automation in understanding conventionally unstructured (raw) multimodal data.

3.1 Event Logs From Multimodal Data

Knoch et al. [12] introduced an unsupervised method for process discovery from video recordings of manual assembly tasks, leveraging overhead cameras to track workers' hands and associate movement patterns with specific work steps, illustrating the potential for practical applications in industrial settings. Kratsch et al. [13] proposed a reference architecture *ViProMiRA* for leveraging video data in process mining, offering a structured approach to transform raw video data into event logs for process analysis, thus expanding the toolkit available for exploring more complex and less structured processes. Lepsien et al. [15] applied process mining to surveillance videos in pigpens, highlighting the importance of further implementation and domain-specific knowledge. They designed an abstract pipeline for process mining on video data, which includes steps from dataset preparation to event log construction and subsequent process mining applications, highlighting the growing stage of this research area. Lepsien et al. [16] used a combination of object tracking, spatio-temporal action detection, and techniques for raising the abstraction level of events and showed the translation of video data into higher-level, discrete event data. Furthermore, Chen et al. [4] concentrated on comparing processes with Petri-net models obtained from videos. The exploration of process mining extends into the realm of sensor data, demonstrating the versatility of process mining techniques in diverse data environments. A significant contribution to this field includes the work by Rebmann et al. [22], who presented a multimodal approach to activity recognition and process discovery, utilizing both motion sensor and video data to enhance the accuracy of captured process activities. Janssen et al. [11] introduced an approach to process model discovery from smart home and IoT sensor event data, showcasing the potential of using sensor activations to map human daily routines through process mining. This methodology divides sensor activation logs into sequential sections, clusters them into patterns of similar sequences, and maps these clusters to activities, which are then grouped into cases based on specific sensor events. Results from a literature review conducted by Telli et

al. [24] show that the term multimodal process mining has been used for multimodal models created by including different perspectives of data in the analysis. In our approach, we use multimodal embedding spaces created by encoder-only architectures, and we compare multimodal data by similarity function (such as cosine similarity).

3.2 Preventing AI Hallucinations

Preventing AI hallucinations in process mining involves ensuring the accuracy and reliability of the extracted process models. Tax et al. [23] suggested using supervised learning for event abstraction, which enhances the quality of process models by reducing noise and improving interpretability. Additionally, semantic-based frameworks like SPMaAF by Okoye et al. [19] have been shown to improve accuracy and conceptual reasoning capabilities in process mining. Folino and Pontieri [6] emphasized the importance of using AI-based strategies to handle low-quality logs, leveraging domain knowledge and auxiliary AI tasks to enhance process mining outcomes. Furthermore, Dixit et al. [5] developed methods for detecting and repairing event ordering imperfections in logs, which is crucial for maintaining high-quality process mining results. These efforts collectively contribute to improving the accuracy and reliability of process mining applications by addressing the challenges associated with AI hallucinations. Our approach deals with AI hallucinations through the human-in-the-loop BICycle method. Figure 2 shows an overview of our solution.

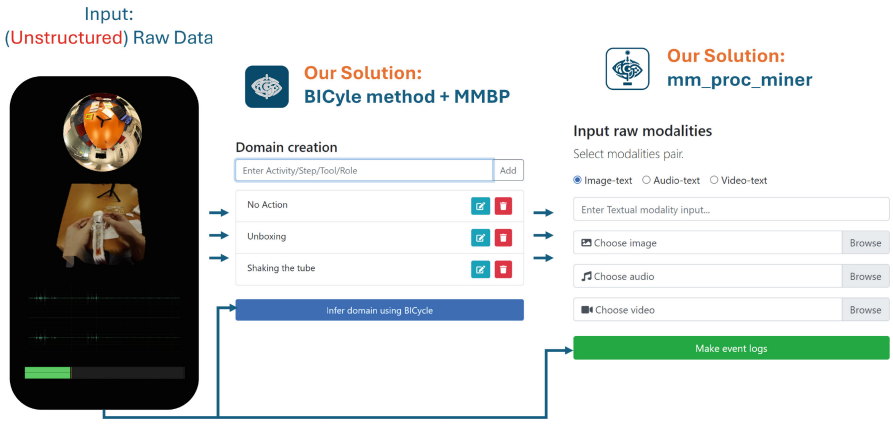

Fig. 2. Overview of our solution. Collected multiple modalities are sent to domain inferring and event log creation (completion).

4 Solution: mm_proc_miner

We have developed an event logs creation tool termed *mm_proc_miner* designed and implemented to enhance the way processes are documented and analyzed

with multimodal awareness. We will begin by introducing essential theoretical foundations, followed by a description of multimodal event logs, and finally, we will discuss the human-in-the-loop method that is used for business-knowledge-aware entity extraction.

4.1 Multimodal Event Logs

To develop multimodal event logs that extend beyond the reach of traditional process mining tools, we explore different data modalities rich in process information yet largely untapped by conventional methods. Video datasets bring the potential for capturing the complexity of human actions in a way that textual data cannot. Videos, with their visual and temporal dimensions, offer a detailed chronicle of processes through the lens of human interactions, both with objects and with one another. Adding depth and realizing relative distances within these interactions becomes possible, further enhancing the comprehension of spatial dynamics critical for understanding complex scenarios.

The auditory modality emerges as another layer of understanding, enriching the insights gained from visual data. The power of sound to convey context, action, and interaction complements the visual narrative, offering a fuller, more nuanced portrayal of events. The significance of integrating auditory data lies in its ability to capture moments and interactions that visual cues alone might not fully reveal. For instance, the sound of a *click* serves as an audible confirmation of actions completed, which can be pivotal in processes requiring precise outcomes. This auditory cue is particularly important in scenarios where visual confirmation may be obstructed or ambiguous.

Consider the example of a home kit for sampling DNA, a process steeped in the need for accuracy and reliability. As users engage with the kit, they follow a sequence of steps, one of which involves closing a funnel lid to secure a sample. The visual action of closing the lid may seem straightforward, yet the auditory *click* sound is what conclusively indicates the lid has been securely fastened as illustrated in Fig. 1. This sound not only assures the user of the successful completion of this step but also serves as an auditory event log, marking a crucial point in the process.

Capturing such auditory cues extends the capability of process mining tools, allowing them to analyze and understand processes that rely on sound as a marker of successful interactions or steps. By integrating auditory data alongside visual recordings, we can develop more comprehensive event logs that capture the full spectrum of human interaction with objects and with one another. This multimodal approach opens up new avenues for analyzing complex processes, enabling the creation of systems that are more responsive and attuned to the details of human behavior.

Building upon the integration of visual, depth, and auditory modalities, the incorporation of sensor data introduces a new dimension to our multimodal event logs, significantly enhancing the traceability and reproducibility of processes such as DNA sampling. Sensors measuring humidity, temperature, and pressure become vital in environments where precise conditions are crucial for

the accuracy and reliability of outcomes. This additional layer of data in our tool can be imported through textual modality to enrich the context around human interactions and actions but also ensure that these processes adhere to the necessary environmental standards.

To evaluate the business-related application of multimodal conformance checking, we have recorded new videos for testing purposes, and we have made the entire dataset publicly available. The videos are out-of-internet wild samples that we guarantee are not used in training any of our base LLM models. We recorded evidence data in two domains. The first domain is DNA sampling in home settings. We recorded two instances of processes in the domain of DNA sample collection using Ancestry DNA [2] test kits and 23andMe [1] DNA test kits (two instances). We show the opportunities of guided and supervised medical applications in home kits that are unlocked with efficient real-time process monitoring from multimodal data. For recording the evaluation video samples, we used the Ray-Ban Meta smart glasses with ultra-wide 12-megapixel cameras for capturing 1080p videos and the Insta360 X3 camera with a 72MP photo resolution, 5.7K 360° video at 30 fps. An overview of the collected dataset is given in Fig. 3.

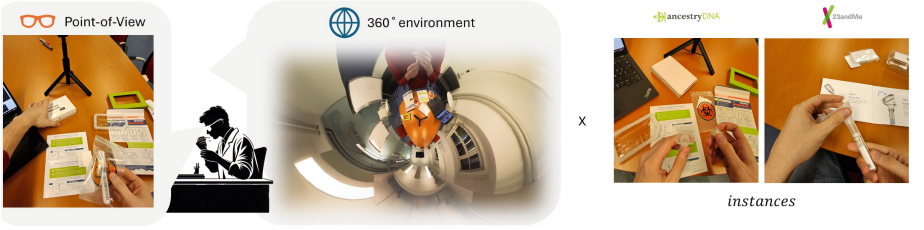

Fig. 3. An overview of the dataset in the domain of DNA home kit collecting that we recorded, processed, and used for testing.

We implemented the integration of multimodal data for the purpose of conversion of raw data into structured insights, employing a pre-trained model, *Imagebind* [9] for joint embedding of diverse data modalities.

4.2 Turning Modalities Into Event Logs

To transform modalities into an event log, we compute the matching scores between the querying modality and keys of collected modalities. Both keys are queries that are simply values of their embeddings, over which we perform similarity matching.

For querying modality, we choose text to keep it compatible with conventional event logs and their textual labels (numeric or characters). We obtain the embedding vector for the querying text using a pre-trained space [9], denoted as E_q, and the embedding vector for the key modality, denoted as E_k. The matching score is calculated by performing a dot product between these embeddings, given

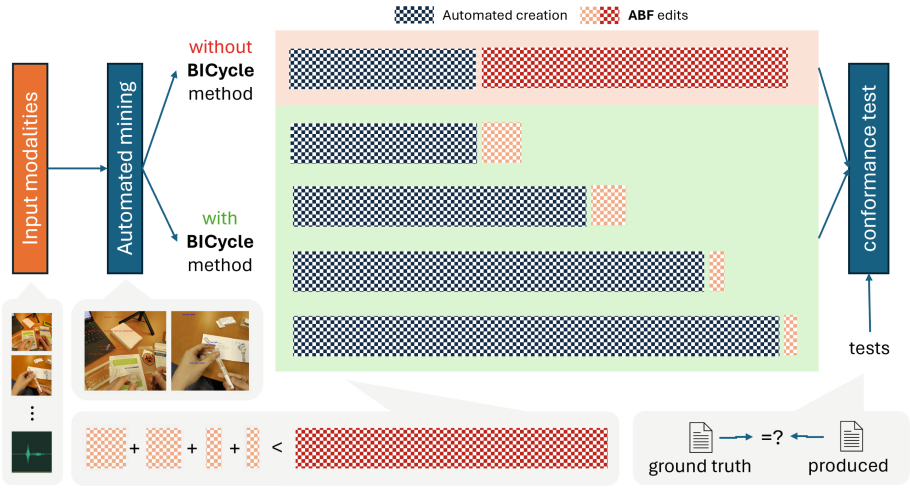

Fig. 4. The BICycle method. ABF - Atomic Business Feedback operations.

by $S = E_q \cdot E_k$. This dot product serves as a similarity measure between the two embeddings. To normalize these similarity scores across multiple key modalities and convert them into a probability distribution, we apply the softmax function. Given multiple key embeddings $\{E_k^1, E_k^2, \ldots, E_k^n\}$, we compute the dot product for each pair, resulting in similarity scores $\{S_1, S_2, \ldots, S_n\}$, where $S_i = E_q \cdot E_k^i$. The softmax function is then applied to these similarity scores to obtain the normalized probabilities P_i for each key modality. This is computed as:

$$P_i = \frac{\exp(S_i)}{\sum_{j=1}^{n} \exp(S_j)}$$

where $\exp(S_i)$ is the exponential of the similarity score S_i, and the denominator is the sum of the exponentials of all similarity scores. This process ensures that the resulting probabilities $\{P_1, P_2, \ldots, P_n\}$ sum to one, providing a robust measure of the likelihood that each key modality matches the querying text. Higher values of P_i indicate a better match between the querying text and the corresponding key modality. Thus, by using the dot product of embeddings and the softmax function, we compute and normalize the matching scores across different modalities, and pick the best match. For instance, if our query is an image modality and the key is text, we can calculate the matching probabilities in PyTorch[2] using:

$$\text{Softmax}(\text{emb}[\text{Modality.VISION}] \cdot \text{emb}[\text{Modality.TEXT}]^{\text{T}},\ \text{dim} = -1)$$

4.3 Business-Knowledge Integration Cycles Method (BICycle)

To facilitate the importance of incorporating domain-specific knowledge into our system, we developed a human-in-the-loop method named the **B**usiness-

[2] https://pytorch.org/.

knowledge **I**ntegration **Cycle**s method (BICycle), which is designed to systematically integrate human (usually a domain expert) feedback into our process mining framework (as illustrated in Fig. 4). This approach ensures that the generated event logs are not only rich in data from various modalities but are also related to relevant concepts of the business domain of the process.

4.4 Preventing AI Hallucinations

At the heart of BICycle are **A**tomic **B**usiness-knowledge **F**eedback (ABF) operations, designed to prevent AI hallucinations (as illustrated in Fig. 5). ABF operations are designed as a granular set of edits that domain experts can perform to refine and enhance the accuracy of the event logs. These operations are crucial for tailoring the automated log generation to the specific contexts and requirements of various fields, making the data more relevant and actionable for users. ABF operations are designed to directly change the embeddings that are keys for the matching described in Sect. 4.2, and include three actions:

- *Instancing.* ABF operation involves specifying a general tool or action mentioned in the event log to a particular named instance. For example, if the automated system identifies a "cutting" action, a domain expert can instance this action to a more specific "cutting with a surgical scalpel" in a medical context. This operation increases the specificity and contextual relevance of the event logs.
- *Renaming.* ABF operation is to fix when terms or labels used by the automated system do not align perfectly with the domain-specific terminology. The renaming operation allows experts to replace these with the correct terminology, enhancing the clarity and usability of the event logs. For example, if the automated system labels an action as "data input," a domain expert might rename this to "patient information entry" in a healthcare setting.
- *Removing.* When certain actions or events captured might be irrelevant or noisy in the context of the specific process being analyzed we can use the *Removing* ABF operation. The removing operation enables experts to delete these entries, ensuring that the event logs remain focused and pertinent. For example, if the event log contains entries for "background noise detection" in an audio processing task, these can be removed to focus solely on relevant audio events.

The BICycle method incorporates these ABF operations into iterative cycles of feedback and refinement. After an initial set of event logs is generated by the system, domain experts review the logs and apply ABF operations where necessary. The system then integrates this feedback, refining its algorithms and improving the accuracy and relevance of the generated logs. This process is not a one-time effort but rather a continuous cycle of improvement. As more feedback is integrated, the system's understanding of domain-specific processes deepens, leading to progressively more accurate and actionable event logs.

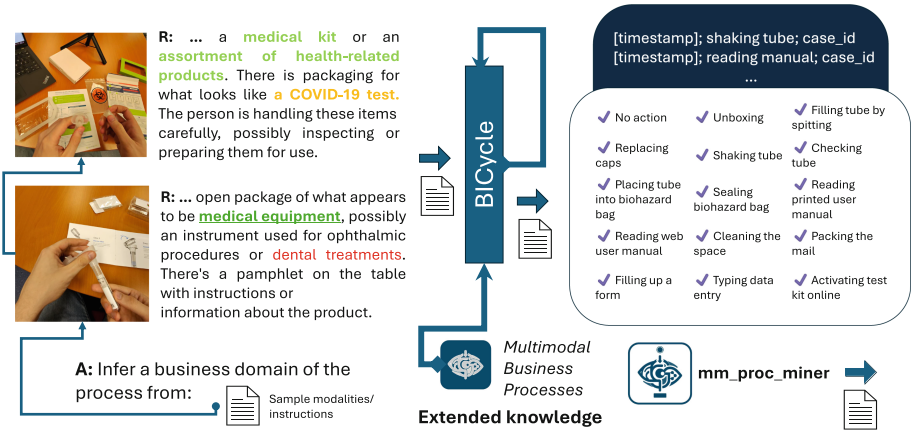

Fig. 5. Preventing AI hallucinations using BICycle.

4.5 Privacy of Evidence Data

Addressing the critical intersection of data privacy and business process management, we propose an **auto-completion** approach to suggest evidence of data missing values within the knowledge about business processes from datasets extracted from natural language processing machine learning models. We introduce the *Multimodal Business Processes dataset*, a comprehensive compilation derived from the concerted efforts of leveraging the state-of-the-art in large language models (LLMs) including Google's Gemini 1.5 [10], and both, OpenAI's GPT4 and GPT3.5 [20].

Our primary objective was to mine and distill extensive business knowledge from vast corpora of multimodal data, predominantly texts and images, across a wide range of business domains. By adopting the teacher-student model, we significantly expanded the knowledge base of a smaller, locally hosted LLaVa 2 [17] multimodal language model.

This approach ensures that sensitive and private data are processed locally, without the need to transmit any information to external cloud servers. This local processing capability is crucial for businesses concerned with data sovereignty and privacy, as it supports secure conformance checking and other data-sensitive operations without requiring an internet connection. An aspect of our solution, particularly relevant to domain inference tasks, involves the selection of a small number of samples from our input domain as initial business knowledge clues. This initial set serves as a foundation upon which we build and refine our understanding of the business domain by employing the BICycle method iteratively. We do this through two steps. The first step is to embed elements of our Multimodal Business Processes dataset into the latent embedding space described in Sect. 4.2. The second step is to include the top K nearest neighboring embeddings in their decoded version (value) as the input to our locally hosted language model.

Table 1. Preview of the Multimodal Business Processes dataset.

Process	Activities
Building a Modular Wine Rack System	Design system, Select wood or metal, Cut and assemble modules, Finish wood or coat metal, Stack and secure modules
Creating Hand-Carved Soap Bars with Custom Scents	Choose soap base, Melt and pour into molds, Carve designs once semi-set, Add essential oils
Making Hand-Tied Bouquets with Dried Flowers	Select dried flowers, Arrange in bouquet, Tie with twine or ribbon, Wrap stems in hessian
Building a Handmade Leather Sling Chair	Design chair, Cut leather for seat, Cut and shape wood for frame, Assemble, Attach leather to frame
Creating Hand-Dipped Beeswax Taper Candles	Melt beeswax, Dip wicks repeatedly until desired thickness, Cool, Trim wicks

4.6 Data Availability

The outcome of our multimodal process auto-completion data is a publicly available dataset comprising 2,644 business processes with a detailed breakdown into 16,180 steps, aimed at enhancing domain inference tasks, as previewed in Table 1.

5 Evaluation

In this section, we evaluate the effectiveness and efficiency of the process model generated from the event logs, which have been enhanced to improve completeness. Our evaluation focuses on three main aspects: *alignment with the official user manual*, *performance and scalability metrics*, and *manual validation of connected modalities*.

5.1 Alignment with User Manual

The process model was meticulously compared with the user manual provided by the DNA test kit company to ensure that it accurately represents the intended operations. This comparison was crucial as the user manual serves as the definitive guide for the process. Through systematic alignment techniques, we confirmed that the model mirrors the step-by-step instructions and sequences described in the manual. This validation ensures that our model not only captures all necessary components of the process but also adheres to the company's operational standards, thereby improving the reliability of the model in practical scenarios.

Using Disco[3] for Process Mining, we analyzed DIY home kits in the domain of DNA collection (Fig. 6). We designed our ground truth in accordance with the provided user manuals by selected AncestryDNA and 23andMe official instructions provided with the kit.

[3] https://fluxicon.com/disco/.

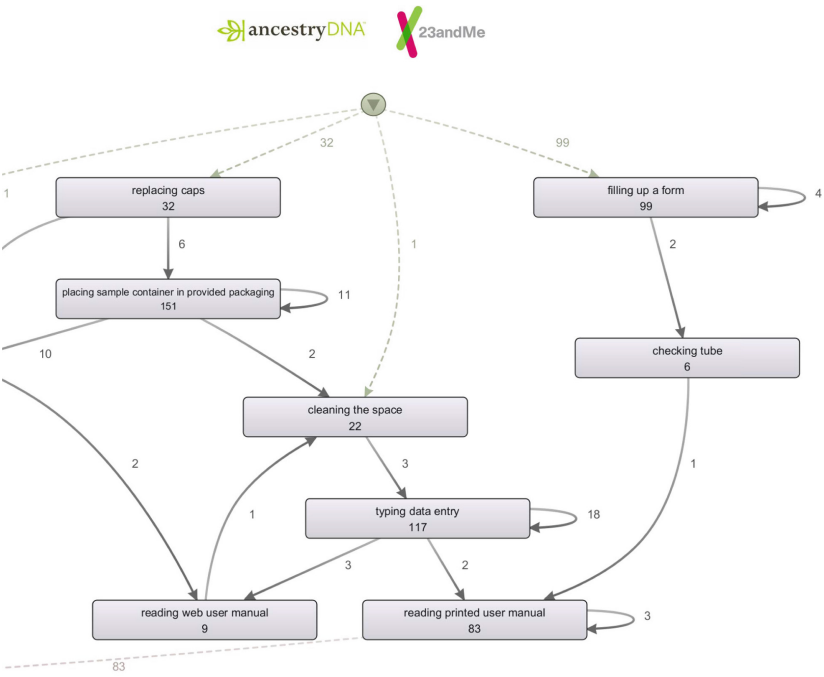

Fig. 6. Process model from our extracted event logs.

5.2 Performance and Scalability

To assess the practical applicability of the enriching of the event logs, we evaluated its performance and scalability on a single A40 GPU with 48GB RAM. Results are given in Fig. 7. The metrics used included process execution time, resource utilization, and scalability under varying loads. The results demonstrate that the model performs efficiently, with low latency, supporting real-time execution. The scalability tests indicate that our model can handle increases in workload without significant degradation in performance, making it suitable for both small-scale and large-scale operations.

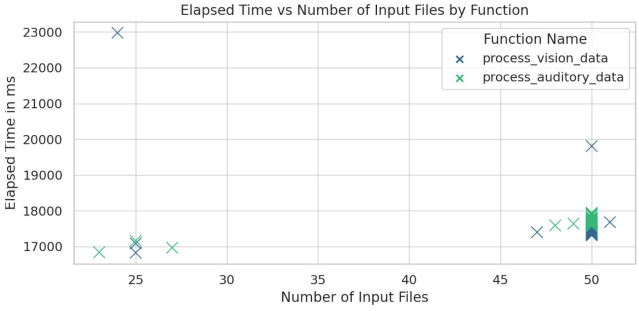

Fig. 7. Evaluation metrics for performance and scalability tests.

5.3 Assessment of the Created Event Logs

Further validation of the event log creation was conducted through manual observations to ensure that all modalities linked to specific steps in the process were accurately represented. Figure 8 provides the results of a manual observation study evaluating the accuracy of event log creation for the DNA testing kits 23andMe and AncestryDNA. Each step of the process was assessed for accuracy in image and audio modalities. These findings confirm the practical accuracy of the model in representing real-life execution, with generally higher accuracy in visual logs compared to audio. This manual check involved scrutinizing the log against observed operations to confirm that each step was correctly associated with its respective modalities. The value of accuracy is calculated over the average of five independent assessments, on the scale from 1.00 to 100.00. This observation confirms the model's practical accuracy in representing the real-life execution of the DNA testing process.

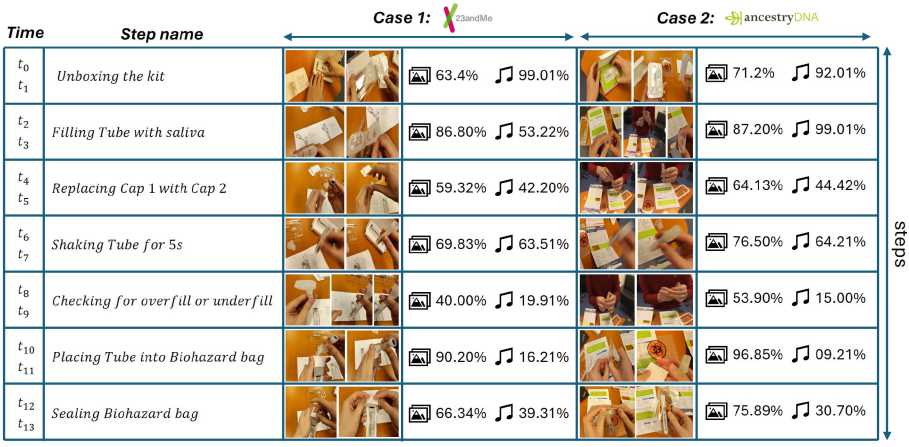

Fig. 8. Validation of created multimodal event log for the test process.

The evaluation of the created event log shows that it is not only a faithful representation of the prescribed process according to the official user manual but also excels in performance and scalability. Additionally, the manual verification of modalities connected to specific process steps further confirms the model's accuracy and utility in real-world applications. This comprehensive evaluation underscores the model's potential to enhance the operational efficiency and reliability of DNA testing processes.

5.4 Evaluating the Human-in-the-Loop Aspect

Table 2 presents the results of evaluating the Human-in-the-loop aspect through ABF operations across ten independent assessments. Without BICycle intervention, the average number of ABFs is 41.2, with individual assessments ranging

Table 2. Evaluating Human-in-the-loop aspect. ABFs over ten independent assessments.

			Assessment No. / Count(ABFs)									
Test		***Avg # of ABFs***	***01***	***02***	***03***	***04***	***05***	***06***	***07***	***08***	***09***	***10***
No-BICycle		41.2	30	26	54	40	48	40	50	46	34	44
BICycle	Iteration 1	3.2	3	1	2	4	1	5	5	5	4	2
	Iteration 2	3	3	3	4	3	5	2	4	2	2	2
	Iteration 3	2.9	1	3	4	2	3	2	5	4	2	3
	Iteration 4	3.2	2	1	5	4	4	4	1	3	3	5
	Iteration 5	2.2	3	1	3	1	4	1	4	2	1	2
	∑	14.5	**12**	**09**	**18**	**14**	**17**	**14**	**19**	**16**	**12**	**14**

from 26 to 54. When BICycle is used, iterative improvements are evident. In Iteration 1, the average number of ABFs drops significantly to 3.2, with assessments ranging from 1 to 5. By Iteration 5, the average further decreases to 2.2, showcasing a reduction in unnecessary or inaccurate entries, with assessments ranging from 1 to 4. The cumulative total of ABFs across all iterations (Iterations 1 to 5) is 14.5, indicating a consistent decrease in ABFs and reflecting the effectiveness of the Human-in-the-loop design in refining and enhancing the event logs.

6 Conclusion

This paper introduced a method for enhancing the fidelity and truthfulness of event log creation through the integration of multiple data modalities and the implementation of a human-in-the-loop business-knowledge integration method. Our method leverages video, audio, text, and sensor data to transform diverse actions, conversations, and interactions into detailed, semantically rich event logs. This multimodal perspective provides a nuanced understanding of business processes, capturing everything from spoken dialogues to physical interactions with unprecedented depth and clarity.

Through the development of the **B**usiness-knowledge **I**ntegration **Cycle**s method (BICycle) and the application of the base set of **A**tomic **B**usiness-knowledge **F**eedback (ABF) operations, we have facilitated a dynamic, iterative process of improvement and refinement. This human-in-the-loop approach ensures that our generative models are continuously informed by domain expertise, enhancing the accuracy and relevance of the generated event logs. Our rigorous formal analyses and testing with out-of-internet samples underscores the effectiveness of our methodology. Data and tools related to our methodology, as well as a test dataset in the domain of DNA collecting home kits, are openly available on our GitHub page, https://github.com/aleksandargavric/mm_proc_miner, ensuring accessibility for further research and application.

References

1. 23andMe (2020). https://www.23andme.com/
2. Ancestry.com: Ancestrydna. Ancestry.com (2020). https://www.ancestry.com/dna/
3. Beerepoot, I., et al.: The biggest business process management problems to solve before we die. Comput. Ind. **146**, 103837 (2023). https://doi.org/10.1016/j.compind.2022.103837
4. Chen, S., Zou, M., Cao, R., Zhao, Z., Zeng, Q.: Video process mining and model matching for intelligent development: conformance checking. Sensors **23**(8), 3812 (2023)
5. Dixit, P.M., et al.: Detection and interactive repair of event ordering imperfection in process logs. In: Krogstie, J., Reijers, H.A. (eds.) CAiSE 2018. LNCS, vol. 10816, pp. 274–290. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91563-0_17
6. Folino, F., Pontieri, L.: Pushing more AI capabilities into process mining to better deal with low-quality logs. In: Di Francescomarino, C., Dijkman, R., Zdun, U. (eds.) BPM 2019. LNBIP, vol. 362, pp. 5–11. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37453-2_1
7. Gavric, A.: Enhancing process understanding through multimodal data analysis and extended reality. In: Companion Proceedings of the 16th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling and the 13th Enterprise Design and Engineering Working Conference (2023)
8. Gavric, A., Bork, D., Proper, H.: Multimodal process mining. In: CBI 2024: 26th International Conference on Business Informatics (2024)
9. Girdhar, R., et al.: ImageBind: one embedding space to bind them all. In: CVPR (2023)
10. Google: Google gemini. Website (2024). https://gemini.google.com. Accessed 2 June 2024
11. Janssen, D., Mannhardt, F., Koschmider, A., van Zelst, S.J.: Process model discovery from sensor event data. In: Leemans, S., Leopold, H. (eds.) ICPM 2020. LNBIP, vol. 406, pp. 69–81. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72693-5_6
12. Knoch, S., Ponpathirkoottam, S., Schwartz, T.: Video-to-model: unsupervised trace extraction from videos for process discovery and conformance checking in manual assembly. In: Fahland, D., Ghidini, C., Becker, J., Dumas, M. (eds.) BPM 2020. LNCS, vol. 12168, pp. 291–308. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58666-9_17
13. Kratsch, W., König, F., Röglinger, M.: Shedding light on blind spots - developing a reference architecture to leverage video data for process mining. Decis. Support Syst. **158**, 113794 (2022). https://doi.org/10.1016/j.dss.2022.113794
14. Körber, N., Wehrli, S., Irrgang, C.: How to measure the intelligence of large language models? (2024). https://arxiv.org/abs/2407.20828
15. Lepsien, A., Bosselmann, J., Melfsen, A., Koschmider, A.: Process mining on video data. In: ZEUS 2022, CEUR Workshop Proceedings, vol. 3113, pp. 56–62. CEUR-WS.org (2022). https://ceur-ws.org/Vol-3113/paper9.pdf
16. Lepsien, A., Koschmider, A., Kratsch, W.: Analytics pipeline for process mining on video data. In: Di Francescomarino, C., Burattin, A., Janiesch, C., Sadiq, S. (eds.) BPM 2023. LNBIP, vol. 490, pp. 196–213. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-41623-1_12

17. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
18. Malinova, M., Gross, S., Mendling, J.: A study into the contingencies of process improvement methods. Inf. Syst. **104**, 101880 (2022). https://doi.org/10.1016/j.is.2021.101880. https://www.sciencedirect.com/science/article/pii/S0306437921001022
19. Okoye, K., Islam, S., Naeem, U., Sharif, M.S., Azam, M.A., Karami, A.: The application of a semantic-based process mining framework on a learning process domain. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) IntelliSys 2018. AISC, vol. 868, pp. 1381–1403. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-01054-6_96
20. OpenAI: ChatGPT (2024). https://chat.openai.com. Accessed 01 Aug 2024
21. Pegoraro, M., van der Aalst, W.M.: Mining uncertain event data in process mining. In: 2019 International Conference on Process Mining (ICPM), pp. 89–96 (2019). https://doi.org/10.1109/ICPM.2019.00023
22. Rebmann, A., Emrich, A., Fettke, P.: Enabling the discovery of manual processes using a multi-modal activity recognition approach. In: Di Francescomarino, C., Dijkman, R., Zdun, U. (eds.) BPM 2019. LNBIP, vol. 362, pp. 130–141. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-37453-2_12
23. Tax, N., Sidorova, N., Haakma, R., van der Aalst, W.M.P.: Event abstraction for process mining using supervised learning techniques. In: Bi, Y., Kapoor, S., Bhatia, R. (eds.) IntelliSys 2016. LNNS, vol. 15, pp. 251–269. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-56994-9_18
24. Telli, A., Erdogan, T.G., Kolukısa, A.: Detecting novel behavior and process enhancement with multimodal process mining. In: 2023 4th International Informatics and Software Engineering Conference (IISEC), pp. 1–6 (2023). https://doi.org/10.1109/IISEC59749.2023.10391012