# Multimodal Process Mining

Aleksandar Gavric
*Business Informatics*
*TU Wien*
Vienna, Austria
aleksandar.gavric@tuwien.ac.at

Dominik Bork
*Business Informatics*
*TU Wien*
Vienna, Austria
dominik.bork@tuwien.ac.at

Henderik Proper
*Business Informatics*
*TU Wien*
Vienna, Austria
henderik.proper@tuwien.ac.at

***Abstract*—Process modeling has a long and successful history, greatly aided by process discovery from event logs. However, many processes leave limited digital traces. Often, manual and physical activities go unrecorded in these logs, leading to significant gaps in the data. To address these blind spots, we must manually model certain activities, which disconnects the conceptual model from actual execution traces. This paper explores how to better track and understand manual business activities that are often poorly documented digitally. We introduce a new method that converts video into detailed event logs, enhancing the way businesses monitor and improve their processes. Our research includes analysis of the *Multimodal Process Mining* concept, and a *Vid2log* tool, designed to improve the way manual activities are recorded and analyzed.**

***Index Terms*—AI-assisted process mining, evidence-based business process management, construction of event logs, multimodal data, world models, AI**

## I. Introduction

Multimodal data, which encompasses diverse types of information such as text, images, audio, and sensory inputs, is crucial for building comprehensive world models [1]. These models are essential not just for human comprehension, but for advancing both non-embodied AI, like virtual assistants, and embodied AI, such as robots navigating real-world environments. This paper connects well established field of process mining [2] with multimodal data. Historically, understanding processes has heavily depended on manual methods such as observation, interpretation, and note-taking. These traditional techniques are not only time-intensive but also susceptible to human errors and inconsistencies. The evolving landscape of process mining, along with the challenges faced by analysts in this field, has been underscored through extensive interviews and surveys like [3], and [4]. At the heart of this journey lies the challenge of capturing, analyzing, and enriching event data for process mining (discovery). The primary objective of process discovery is to generate a process model that accurately reflects the underlying process, based on an analysis of event logs containing examples of behaviors [5]. We focus the research question of this paper on: *How can information systems become more aware of real-life processes characterized by manual steps and limited digital traceability by enriching event logs with additional modalities such as video data?*

We address this challenge by introducing a *Vid2log* tool, to bridge the gap between multimodal data sources and the creation of comprehensive process-related event logs.

Following a study of a general model for information coverage [6], which shows that advisory systems should not only provide relevant documents to searchers but also help them effectively cover their information needs, we are exploring the knowledge gains of enriched event logs. We aim to explore the critical dimensions of *who, when, where, how,* and *why* in each process step to assess the tool's effectiveness in transforming event logs into more detailed and actionable insights. In Section II, we define our perspective on *Multimodal Process Mining*. Section III provides an overview of our developed solution, Vid2log. Section IV presents the evaluation of our solution. After specifying our concept and the scope of our research, we discuss related work in Section V. We conclude in Section VI.

## II. Multimodal Process Mining

**Multimodal Process Mining (MMPM)** is a method to analyze business processes by combining different types of data. This method uses visual, auditory, sensor, and machine log observations (*modalities*) to understand different aspects of activities within a business. By bringing together these different data sources, MMPM offers a complete view that traditional single-modality (i.e. just text) analyses might miss.

Figure 1 shows how these different types of observations connect to a central business process. Visual modality observations include statics (such as *objects and tools*), dynamics (such as *motions*), relations (such as *the act of someone dragging something*), and environmental (such as *a scene set in an office environment*). Auditory modality observations cover sounds that can trigger some action (such as *applause, a siren, or a speech that has just started*), are continuously present (such as *the sound of an engine running*), are countable (such as *hit tests or sound events*), or focused on the delivery of sound itself (such as *from which spatial direction or in what pitch*). Unstructured sensor observations gather raw data on factors like temperature and humidity, while machine-output logs include both manually entered data, such as *customer details*, and automatically recorded information, such as *usage metrics*. Together, these observations create a detailed picture of the business process, showing both small details and the larger context of *Process Mining*.

Process mining [2] is a technique used to discover, monitor, and improve business processes by extracting knowledge

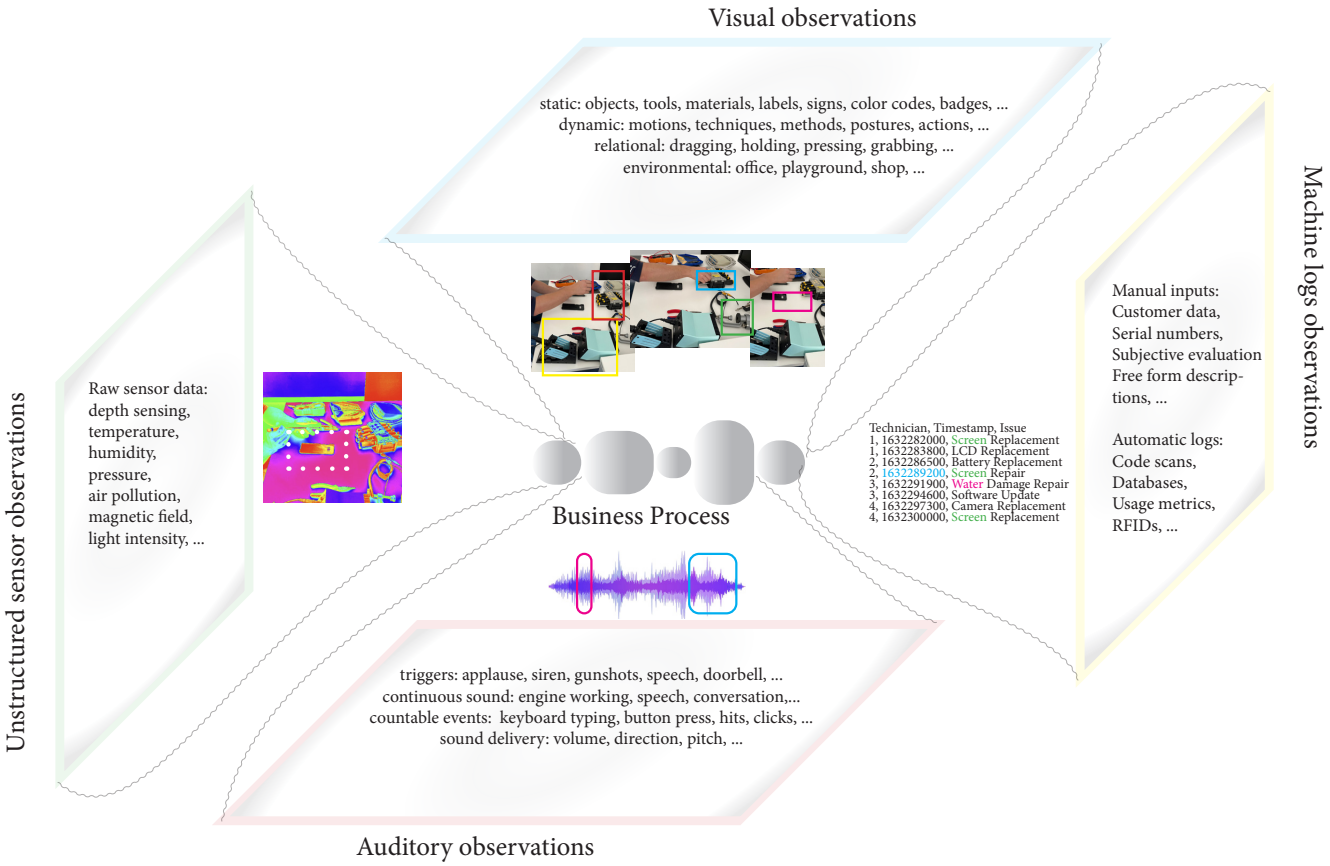


Fig. 1. Multimodal traces of a business process [7]

from event logs commonly available in today's information systems. According to [8], process mining techniques help organizations discover and analyze business processes based on raw event data, emphasizing the context in which events occur. Van der Aalst [9] further describes process mining as a bridge between data mining and business process modeling, emphasizing its role in analyzing end-to-end processes due to the growing availability of event data and new discovery and conformance checking techniques. Srivastava et al. [10] highlight process mining as an interdisciplinary domain that encompasses process modeling, analysis, and business intelligence, demonstrating its utility in generating faster results and ensuring process conformance and compliance. These foundational insights into process mining set the stage for understanding multimodal process mining, which extends these principles by integrating data from multiple data modalities to provide a richer, more detailed view of business processes that can be generated with automation.

*Definition 1 (Multimodal Process Mining (MMPM)):* Let $M = \{m_1, m_2, \ldots, m_k\}$ represent the set of modalities. The comprehensive event log $E$ is constructed as $E = \bigcup_{i=1}^{k} E_i$, where $E_i = \{(e_{i1}, t_{i1}, a_{i1}), (e_{i2}, t_{i2}, a_{i2}), \ldots, (e_{in}, t_{in}, a_{in})\}$ captures the events, timestamps, and attributes for each modality $m_i$. The event log $E$ integrates these individual logs as $E = \bigcup_{i=1}^{k} \{(e, t, a) \mid (e, t, a) \in E_i\}$. The discovery function, defined as $\mathcal{P} = \text{Discover}(E)$, applies algorithms to analyze $E$ and create a process model $\mathcal{P}$. Thus, MMPM captures multimodal contributions to knowledge gain in process mining, and is defined as

$$\text{MMPM} = \left( M, E = \bigcup_{i=1}^{k} E_i, \mathcal{P} = \text{Discover}(E) \right).$$

The enriched event log $E$, as defined in MMPM, serves as the foundational input for **Multimodal Conformance/Compliance Analysis (MMCCA)**. By utilizing the detailed event logs generated through MMPM, MMCCA applies conformance functions to evaluate how well the observed behavior aligns with predefined process models $P$ and complies with specified rules R. The conformance score $C$ derived from MMCCA provides a quantitative measure of this alignment and compliance. Hence, MMPM facilitates the collection and integration of diverse observational data, while MMCCA uses this enriched data to ensure that business processes not only adhere to intended models but also meet regulatory and organizational standards, thus enhancing process accuracy and reliability.

*Definition 2 (Multimodal Conformance/Compliance Analysis (MMCCA)):* Let $M = \{m_1, m_2, \ldots, m_k\}$ be the set of modalities. For each modality $m_i$, let $E_i = \{(e_{i1}, t_{i1}, a_{i1}), (e_{i2}, t_{i2}, a_{i2}), \ldots, (e_{in}, t_{in}, a_{in})\}$ represent the event log. The comprehensive event log $E$ is constructed as $E = \bigcup_{i=1}^{k} E_i$. Given a process model $P$ and a set of conformance rules R, the conformance function $\text{Conform}(E, P, R)$ evaluates the degree to which the observed behavior $E$ adheres to the process model $P$ and complies with the rules R. Define $C$ as the conformance score: $C = \text{Conform}(E, P, R)$. Thus, Multimodal Conformance/Compliance Analysis is defined as:

$$\text{MMCCA} = \left( M, E = \bigcup_{i=1}^{k} E_i, P, R, C = \text{Conform}(E, P, R) \right)$$

where $M$ is the set of modalities, $E$ is the enriched event log, $P$ is the process model, R is the set of conformance rules, and $C$ is the conformance score, representing the alignment and compliance of the observed behavior with the predefined models and rules.

## III. *Vid2log* TOOL

We have developed a *Vid2log* tool, detailed in Fig. 2, designed and implemented to incorporate multimodal data in the way processes are documented and analyzed. Through large language models (LLMs), we use a method called tokenization to break down the input data into smaller, manageable pieces, which allows us to use the pre-trained capabilities of Multimodal LLMs to process and analyze multimodal data effectively. Tokenization is the process of converting a sequence of characters into a sequence of tokens (e.g., words, subwords, or characters). Multimodal LLMs take the concept of tokenization and apply it to different types (modalities) of data. For simplicity, assume we have a vocabulary $V$. For an example, given an input text $\mathbf{T}$ = "Process is finished!", the tokenization process maps this text into a sequence of tokens (assumable): $\mathbf{T}$ = ["Process", "is", "finished", "!"]. Each token is then mapped to an integer ID using the vocabulary $V$: $\mathbf{x} = [x_1, x_2, x_3, x_4]$, where $x_i \in \mathbb{N}$ and corresponds to the index of the token in the vocabulary.

These token IDs are converted into dense vectors using an embedding matrix $\mathbf{E} \in \mathbb{R}^{|V| \times d}$, where $d$ is the embedding dimension. The embedding for each token is $\mathbf{e}_i = \mathbf{E}[x_i]$. The input sequence is then represented as $\mathbf{X} = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4]$.

The machine learning architecture that is used for building the predictive model is a *transformer architecture* which primarily consists of self-attention mechanisms and feed-forward neural networks. Self-attention computes a representation of each token in the context of all other tokens in the sequence. Given the input matrix $\mathbf{X}$, self-attention works as follows:

(1) *Linear Transformations*: Compute the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ matrices: $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$, and $\mathbf{V} = \mathbf{X}\mathbf{W}^V$, where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d_k}$.

(2) *Scaled Dot-Product Attention*: Compute the attention scores:

$$\mathbf{A} = \text{softmax}\left( \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} \right).$$

Lastly, apply the attention scores to the values: $\mathbf{Z} = \mathbf{A}\mathbf{V}$.

(3) *Concatenation and Linear Transformation*: If using multi-head attention, concatenate the outputs from multiple attention heads and apply a final linear transformation: $\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_h)\mathbf{W}^O$, where $\mathbf{W}^O \in \mathbb{R}^{hd_k \times d}$ and $h$ is the number of attention heads. After the attention mechanism, the output is passed through a feed-forward neural network, applied independently to each position: $\text{FFN}(\mathbf{z}_i) = \max(0, \mathbf{z}_i\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$.

A single transformer layer consists of the following components: *Self-Attention Mechanism*, *Add & Norm* (incorporating residual connection and layer normalization), *Feed-Forward Network (FFN)*, and *Add & Norm* (incorporating residual connection and layer normalization). The complete transformer model is formed by stacking multiple transformer layers. The output from the final transformer layer is used for downstream tasks, such as classification or sequence generation. For language modeling, the final output is often a probability distribution over the vocabulary, which is computed using a softmax function applied to the linear transformation of the final hidden states: $\mathbf{y}_i = \text{softmax}(\mathbf{h}_i\mathbf{W}^P + \mathbf{b}^P)$, where $\mathbf{W}^P \in \mathbb{R}^{d \times |V|}$ and $\mathbf{b}^P \in \mathbb{R}^{|V|}$.

In this paper, we focus on using LLM (in particualr, LLaVA [11]) to tokenize and analyze text and images, as a proof-of-concept for benefits of multimodal process mining. We start with a clustering algorithm to remove similar frames, reducing redundancy and making the analysis more efficient. After filtering redundant frames, we select the most informative frames from video data using a custom logic that focuses on identifying blind spots, especially in activities that are very manual or physical. This ensures that we choose frames rich in relevant information for the business process we are studying. Then our solution employs a Language Model (LLaVA) to infer the domain of the process. Leveraging this inferred domain knowledge, the system then utilizes instructions aimed at maximizing *Knowledge Gain* that we will define as follows. Our solution enriches both, videos and conventional event logs, with responses to event-related questions like *who, when, where, how, and why*. Subsequently, these partial event logs, derived from various segments of the video and conventional logs, are combined into a final, unified event log. This amalgamation is achieved through a process of *Temporal semantic matching*, where the system aligns and merges logs based on their occurrences in time and contextual relations.

### A. Video Frame Selection

First step in focusing on business-relevant parts of video is done through a reduction of the number of frames to be considered from the video, for which we designed a clustering-based selection.

Let $N$ be the total number of frames available, and let $K$ represent the number of clusters to be formed. We define $C_i$ as the $i$-th cluster, where $i$ ranges from 1 to $K$, and let $F_k$ denote the selected frame from cluster $C_k$, where $k$ ranges from 1 to $K$.

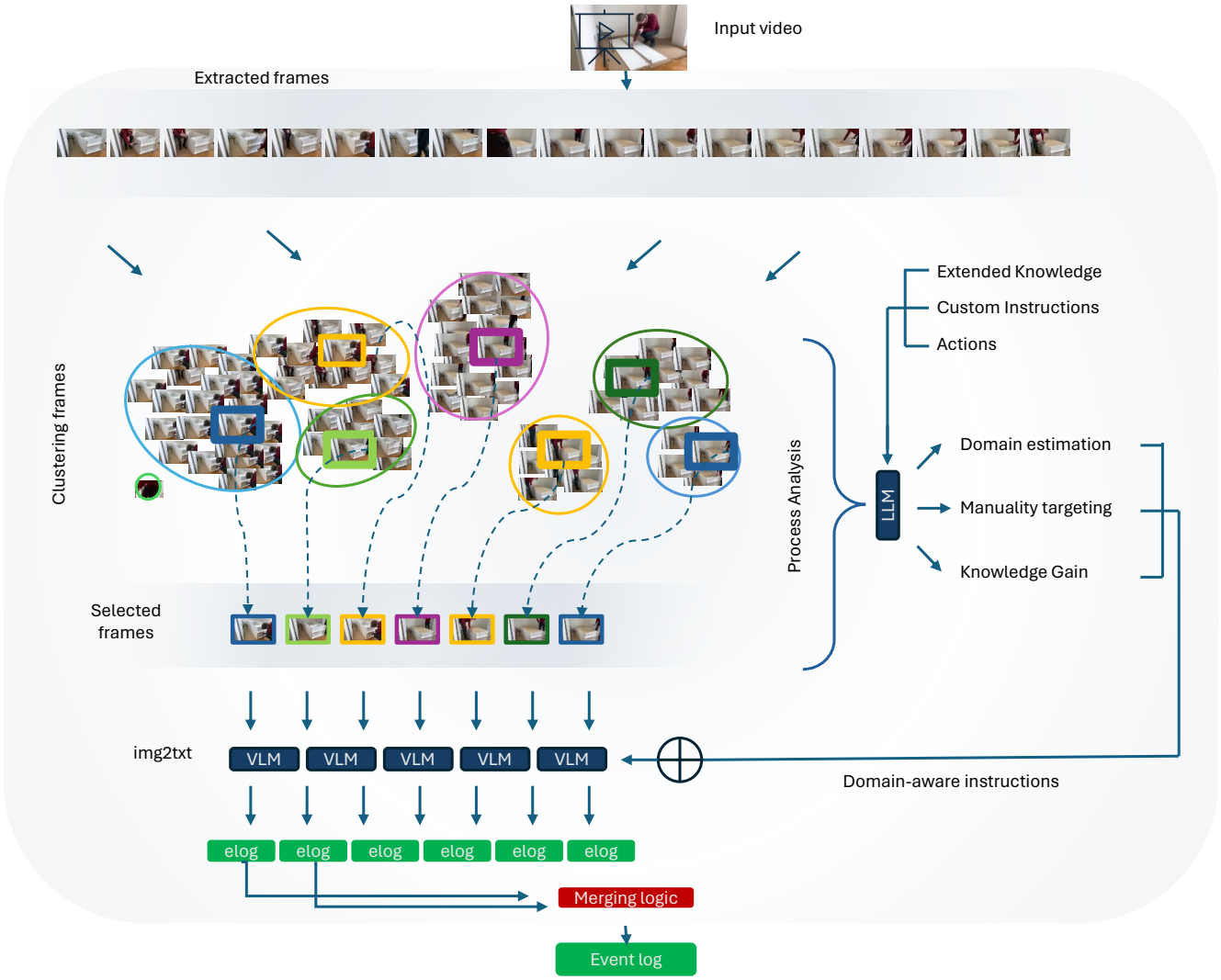


Fig. 2. An illustration of our *Vid2log* tool

The clustering-based selection process consists of the following steps: (1) First, clustering using K-Means, where the $N$ frames are clustered into $K$ clusters using the K-Means algorithm: $C_1, C_2, \ldots, C_K$. (2) Then, selecting centroids as selected frames, where the centroid of each cluster is chosen as a selected frame: $F = \{\text{Centroid}(C_1), \ldots, \text{Centroid}(C_K)$.

Our objective function for clustering and centroid selection is thus:

$$G(F) = \sum_{i=1}^{K} \sum_{j \in C_i} \text{similarity}(\text{Centroid}(C_i), \text{Frame}_j).$$

This objective function calculates sum of similarities between each selected centroid and all frames within the cluster it represents.

### B. Search for Business-Relevant Activities

We use the concept of *Domain Estimation* and *Manuality* as a way of communication with an LLM to narrow the scope of all activities and resources present in raw data to a set of business-relevant process evidences (see Fig. 2).

Firstly we prompt LLM to estimate business domain of our data. Then we retrieve a set of business-relevant activities and resources for the estimated domain, again, by prompting LLM. Our search for business-relevant activities continues through prompts about activites with high Manuality. Manuality is a metric value that measures the level of difficulty in digitally tracking a process without employing multimodal analysis techniques. Manuality seeks to quantify the extent to which a process necessitates the integration of multiple data streams or modes to gain a comprehensive understanding.

The concept of Manuality stems from the recognition that many real-world processes involve complex interactions that cannot be fully captured by single-dimensional data sources. For instance, tracking the quality of a manufacturing process might require the simultaneous analysis of sensor data, visual inspection, and audio feedback. In such cases, the level of

Manuality would be high, indicating a significant reliance on multimodal analysis to monitor and control the process effectively. Processes with high Manuality values require more sophisticated data collection methods and technology investments to ensure accurate monitoring and control. Conversely, processes with low Manuality values are more amenable to straightforward, unimodal tracking. Understanding the Manuality of different processes within an organization can guide technology adoption strategies and help prioritize data analytics initiatives.

### C. Image-to-Text Conversion

Once the clustering step is finished and the business-relevance filters are applied, we use extracted evidence representatives as a prompt to LLM (LLaVA [12]) to perform *Image-to-text conversion* (img2text) in accordance to the evidence-relevant information (*who, when, where, how, and why*) for each step of a process.

Firstly, the 'who' aspect of prompting focuses on identifying actors involved in each step of the process. By querying the LLM about the individuals or entities participating in a step, we gain insights into roles, responsibilities, and interactions. This aspect of information is used for understanding the human or systemic agents driving the process. Secondly, addressing the 'when' component involves determining the timing and sequence of each process step. Prompting the LLM to extract temporal data ensures a chronological understanding of the process flow. This is used for analyzing the timeline of the process. By identifying when each step occurs, the LLM helps in constructing a temporal map of the process, which is essential for optimization and monitoring. We structure this temporal information not to be defined by a timestamp (as a timestamp is already available from the frame position in the video) but to be relational, giving an answer to more specific question - *after what* and *before what* this particular activity happens. The 'where' prompt searches into the spatial context of each process step. Understanding the location or environment where each step occurs is used for assessing resource allocation, environmental constraints, and logistical considerations. The 'how' aspect is concerned with the methods, tools, and procedures employed in each step. By prompting the LLM to analyze the modes of operation, techniques, and resources utilized, we obtain a detailed understanding of the execution of each step. Lastly, the 'why' prompt aims to uncover the rationale or purpose behind each step. Understanding the objectives and reasons for each action provides a strategic perspective of the process. By prompting the LLM to analyze the underlying motivations and goals, we gain insights into the process's relevance.

### D. Knowledge Gain

Another step of structuring evidence data into a more business-relevant event log representation involves filtering activities and resources based on the knowledge they can contribute if they are included in the log. We perform this using a *Knowledge Gain* filter. Knowledge Gain is the potential for acquiring additional information, insights, or data about a particular process or activity by incorporating various digital technologies and data sources. High Knowledge Gain potential are associated with processes that have a high level of Manuality. Knowledge Gain considers the increase in information or data obtained by incorporating additional digital tools or data sources.

Let $K_{\text{initial}}$ represent the initial level of knowledge or information available about a process through traditional IT system interactions or event logs. Let $K_{\text{enhanced}}$ represent the increased level of knowledge or information gained by incorporating additional digital technologies, such as cameras, sensors, and other data sources. Then, Knowledge Gain ($KG$) are represented as: $KG = K_{\text{enhanced}} - K_{\text{initial}}$.

The difference between $K_{\text{enhanced}}$ and $K_{\text{initial}}$ quantifies the increase in knowledge due to the incorporation of digital technologies and additional data sources. We can represent Knowledge Gain using equations based on entropy, which is often used to measure uncertainty or information content. Let $P_{\text{initial}}$ be the probability distribution of events or states before incorporating additional data sources or digital tools. Let $P_{\text{enhanced}}$ be the probability distribution of events or states after incorporating additional data sources or digital tools. The entropy $H$ is then defined as:

$$H(P) = -\sum [p(x) \log_2(p(x))]$$

where: $H(P)$ is the entropy of the probability distribution $P$, and $p(x)$ represents the probability of a specific event or state $x$ in the distribution.

Then, Knowledge Gain (KG) are represented as the difference in entropy: $KG = H(P_{\text{initial}}) - H(P_{\text{enhanced}})$. We can prompt LLM to estimate $P_{\text{initial}}$, while $P_{\text{enhanced}}$ we can calculate from the video.

### E. Merging Logic for Unifying Multimodal Event Logs

The merging logic for unifying multimodal event logs involves methods to integrate information from various event-log sources while maintaining semantic coherence and temporal accuracy. Temporal stamps (available directly from videos) play a critical role in this process, as they allow for the chronological ordering of events. By aligning events based on their timestamps, the merging logic ensures that the sequence of occurrences is accurately preserved. This temporal alignment is essential for understanding the flow of activities and their interdependencies over time, which is crucial for subsequent analysis and decision-making on case identifiers (IDs).

Beyond merely comparing timestamps, the merging logic also examines semantic relationships between concepts. This involves analyzing the content and context of events to identify meaningful connections and associations. For instance, events that describe similar or related activities, even if they occur at different times, are grouped together to provide a more comprehensive understanding of a process. This semantic analysis helps in clustering related events, thereby enhancing the interpretability of the event logs. It also aids in detecting

patterns and trends that might not be evident through temporal analysis alone.

After ordering events and identifying semantic relationships, the merging logic estimates case IDs based on resource usage. Resources, such as personnel or equipment involved in specific events, provide clues for linking related events. By tracking the involvement of resources across different events, the merging logic infers which events belong to the same case or process instance. This estimation of case IDs ensures that events are correctly attributed to their respective cases, facilitating accurate process analysis and reporting.

## IV. Evaluation

We firstly propose evaluation perspective for MMPM, through a comparison of event log before and after the application of MMPM. Secondly, in aims to quantitatively assess the efficacy of our AI-enabled event logs creation pipeline and *Vid2log* tool, we designed a case study.

### A. Evaluation Perspective for Multimodal Process Mining

We propose the application of a *Richness* (R) function in evaluating the completeness of multimodal process mining-related event logs. This metric transcends mere data quantity, emphasizing the quality and relevance of information captured in each step of the process.

Multimodal process mining-related completeness in event logs is a degree to which a log encapsulates detailed and comprehensive information about a process. This is captured through a metric called *Step-wise Richness* (R). For a given event log L, R(L) quantifies the extent of information coverage regarding key questions: who, when, where, how, and why for each step. Step-wise Richness (R) of a process is a metric designed to quantitatively evaluate the depth and comprehensiveness of information captured in an event log for each step of a business process. It quantifies how well an event log answers key questions about each step of the process: who is involved, when it occurs, where it takes place, how it is performed, and why it is necessary. For each step in the process, $R$ is calculated by assessing the quality of information available in the event log regarding the aforementioned dimensions. This involve a domain-aware scoring system where points are allocated for each dimension based on the detail and accuracy of the recorded information. The success is achieved if, in the comparison of the step-wise richness of an event log before ($R(L)$) and after ($R(L')$) application of MMPM, it is true that $R(L) \leq R(L')$.

This inequality asserts that the multimodal enrichment enhances the completeness of the event log.

#### *Enrichment of the Event Log*

The enrichment ($E$) is defined as the difference in richness due to the tool: $E = R(L') - R(L)$.

$E$ varies within the range [0, $E_{\text{max}}$], where $E_{\text{max}}$ is defined as $\frac{1}{\text{Manuality}}$. The Manuality metric, as previously defined, inversely correlates with the potential for completeness increase.

For a process with $N$ steps, the Total Enrichment ($E_{\text{total}}$) is the summation of individual enrichments for each step:

$$E_{\text{total}} = \sum_{i=0}^{N} E(i).$$

Weighted Total Enrichment incorporates the relative importance of each step, $E_{\text{total\_weighted}} = \sum_{i=0}^{N} E(i) \times w_i$, where $w_i$ is the weight assigned to the $i$th step.

#### *Relational Enrichment ($E_{rel}$)*

The Relational Enrichment ($E_{\text{rel}}$) matrix is a construct designed to capture and quantify the indirect enrichment of each step in a process when another step is enriched through the use of MMPM. This matrix helps in understanding the interconnectedness of different steps and their cumulative impact on the overall process.

Simply represented, $E_{\text{rel}}$ is defined as a matrix representing how enrichment of one step indirectly affects others, meaning, for steps $i$ and $j$, $E_{\text{rel}}[i][j] =$ Indirect Enrichment of step $i$ due to enrichment of step $j$.

The diagonal elements are 1, representing direct enrichment. Weighted relational enrichment is calculated by applying weights to the inter-step relationships.

$E_{\text{rel}}[i][j]$ is calculated based on how the enrichment of step $S_j$ (through the application of MMPM) indirectly enhances the completeness of step $S_i$. If the enrichment of step $S_j$ has no impact on step $S_i$, $E_{\text{rel}}[i][j] = 0$. If the enrichment of step $S_j$ directly impacts step $S_i$, then $E_{\text{rel}}[i][j]$ is a positive value, which are determined based on the degree of this impact.

The complete $E_{\text{rel}}$ matrix is constructed by determining the impact values for each pair of steps in the process. The matrix essentially represents a directed graph, where each node corresponds to a process step, and the edges (with weights) represent the influence of one step's enrichment on another. The determinant of the $E_{\text{rel}}$ matrix, denoted as $\det(E_{\text{rel}})$, are calculated to provide a singular value representing the overall relational enrichment of the process. Additionally, the matrix are analyzed to identify key steps that have the most significant influence on others, guiding process optimization efforts.

To define a weighted version of the $E_{\text{rel}}$ matrix, we need to incorporate additional weights that reflect the relative importance or influence of each process step on the others. These weights are based on various factors like the criticality of the steps, their centrality in the process, or their impact on the overall outcome. Let's denote these weights as $w_i$ for step $S_i$. The weighted $E_{\text{rel}}$ matrix are expressed as follows. Assign a weight $w_i$ to each step $S_i$ in the process. This weight represents the relative importance or influence of step $S_i$ in the overall process. The weighted impact of enriching step $S_j$ on step $S_i$ is then given by $E\_\text{rel\_weighted}[i][j] = w\_i \cdot E\_\text{rel}[i][j]$. Here, $E_{\text{rel}}[i][j]$ is the original enrichment impact from step $S_j$ to step $S_i$, as defined in the $E_{\text{rel}}$ matrix. The complete weighted $E_{\text{rel}}$ matrix is constructed by calculating $E_{\text{rel\_weighted}}[i][j]$ for each pair of steps ($S_i$, $S_j$) in the process. If the process has N steps, the weighted $E_{\text{rel}}$ matrix will be an NxN matrix, similar to the original $E_{\text{rel}}$ matrix, but with each element adjusted by the corresponding weight, as represented in (1).

$$E_{\text{rel\_weighted}} = \begin{bmatrix} w_1 \cdot E_{\text{rel}}[1][1] & w_1 \cdot E_{\text{rel}}[1][2] & \cdots & w_1 \cdot E_{\text{rel}}[1][N] \\ w_2 \cdot E_{\text{rel}}[2][1] & w_2 \cdot E_{\text{rel}}[2][2] & \cdots & w_2 \cdot E_{\text{rel}}[2][N] \\ \vdots & \vdots & \ddots & \vdots \\ w_N \cdot E_{\text{rel}}[N][1] & w_N \cdot E_{\text{rel}}[N][2] & \cdots & w_N \cdot E_{\text{rel}}[N][N] \end{bmatrix} \quad (1)$$

### B. Real World Case Study

In our pursuit to demonstrate the practical application and efficacy of our *Vid2Log* tool, we conducted a comprehensive case study centered around the assembly of an IKEA wardrobe. This seemingly manual task, often riddled with complexities and decision-making junctures, served as an ideal candidate for our evaluation. We recorded the assembly process (according to the IKEA's user manual [13]), spanning over multiple consequent videos (as illustrated in Fig. 3). This video process evidence provided us with a rich source of observational data for MMPM, encapsulating various steps, actions, and decisions inherent in the furniture assembly process. *Vid2log* tool and evaluation data are available on our GitHub page[1].

Fig. 3. Preview of the real world video data used for evaluation.

Upon generating the process model (given in Fig. 4) using DISCO [14], we proceeded to check for compliance with the guidelines specified in the IKEA user manual [13]. The (MMCCA) compliance checking followed the user manual's stipulations to ensure that steps and actions in the generated process model adhered to the prescribed instructions. We confirmed that both the user manual and the generated process model were in alignment, validating the accuracy and reliability of the *Vid2Log* tool in capturing and modeling the real-world assembly process.

To extend the utility of MMPM and *Vid2Log* tool beyond the domain of real world evidence recordings, we explored its potential to the scope of business process simulation. This shift required a transition from real-world images to AI-generated business process images using [15], transforming the physical assembly scenarios into a digital environment that mimics a video-game-like simulation (as illustrated in Fig. 5). This approach allowed us to recreate complex business processes in a controlled, interactive setting, where each step could be visualized and analyzed with greater clarity and detail.

We undertook this transition to address scenarios that are challenging to explain through words alone but are equally difficult to record due to resource constraints or regulatory limitations (such as GDPR [16] or privacy). By leveraging AI techniques to generate realistic and dynamic business process images, we were able to emulate decision points, workflows, and interactions typical of business operations. These AI-generated simulations allowed us to illustrate and explore complex business scenarios that would otherwise be infeasible to document, while not relaying on the requirement that processes needs to be described only with words (single modality).

This approach enabled us to include event log evidence that aligns with the extended domain logic of the process, as illustrated in the activity diagram shown in Fig. 6 - extending from furniture assembly to furniture production. By simulating business processes through AI-generated simulations, we showed the ability to capture and document events that are integral to the process flow but potentially challenging to represent in real-world recordings.

In terms of scalability and performance, our *Vid2Log* tool demonstrates efficiency and reliability. The tool is capable of operating in real-time, with its performance scaling linearly with the length of the video. This ensures that even lengthy and complex process recordings can be processed. To address privacy concerns, we designed our solution for self-hosting of all AI models, ensuring that no data is uploaded to third-party servers, thus maintaining strict control over sensitive information. Additionally, to mitigate the risk of AI hallucinations, we empower human moderators to configure filtering parameters using our concepts of Domain Estimation, Manuality, and Knowledge Gain, and enhance the precision and trustworthiness of the generated process models, combining the strengths of AI with the critical oversight of human expertise.

## V. Related Work

In the field of process mining, various studies have evaluated the use of event logs to improve process understanding and efficiency. Adriansyah and Buijs (2012) analyzed event logs from a Dutch financial institute to uncover process performance insights and deviations using alignment techniques [17]. Sonawane and Patki (2015) proposed an automatic system for generating process models from unstructured event logs [18]. Martin, Pufahl, and Mannhardt (2021) developed

[1] Vid2log and evaluation data: https://github.com/aleksandargavric/vid2log

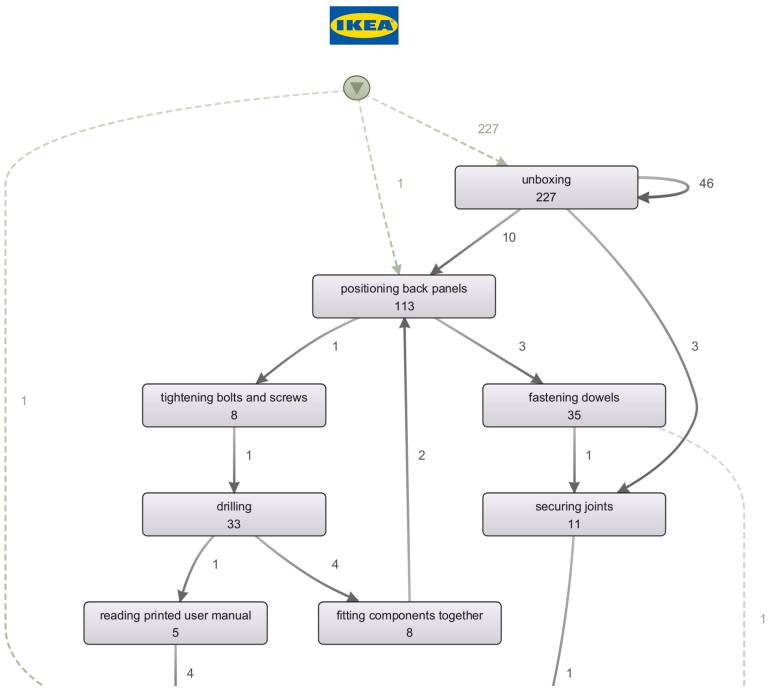


Fig. 4. Process model as a result of process discovery from the real world video data. (Example-domain labels intentionally left unreadable)

Fig. 5. Extended evidence data set using generated data

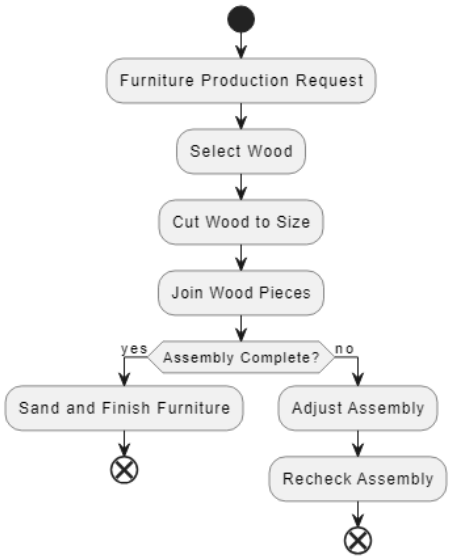


Fig. 6. Extended evidence data set using generated data

an algorithm to detect batch processing in subprocesses from event logs [19]. Pegoraro and van der Aalst (2019) explored process discovery and conformance checking challenges with uncertain event logs [20]. Dixit et al. (2018) introduced techniques for detecting and repairing event ordering issues in event logs [21]. Song et al. (2017) presented an efficient method to align event logs with process models, reducing search space for optimal alignment [22]. Utama et al. (2020) proposed a method to incorporate shift work information into simulation models from event logs using clustering techniques [23]. Marin-Castro and Tello-Leal (2021) reviewed event data preprocessing techniques and their impact on process mining tasks [24]. Song et al. (2015) proposed a heuristic approach for recovering missing events in process logs using process decomposition [25]. Our contribution is in the domain of extending the scope of types of data that can be used in automation to enable all of the aforementioned applications.

Through our WH-questions structuring and Knowledge Gain metric, we extend the concept of *aboutness* [26] that refers to the relationship between different information carriers, where one carrier (a ”promise”) defines what another

carrier is about. Various mechanisms, such as keywords, vectors, or conceptual graphs, have been developed to characterize the aboutness of information carriers. In a theory of demand and supply from [6] authors examine influence between information carriers. We show progress in equipping advisory systems with reasoning abilities to guide creating digital traces of broad spectrum of processes.

Incorporation of multimodal data that we propese in MMPM concept and in our Vid2log tool is connecting data science with process science. Therefore, we structure the rest of our related work section into two folds: (1) data mining on multimodal data and (2) process mining on multimodal data.

### A. Data mining on multimodal data

Tosi et al. [27] introduced a novel framework that aims for comprehensive scene understanding from videos by learning depth, motion, and semantics simultaneously from monocular videos. This method utilizes knowledge distillation and self-supervision to create a compact network architecture, enabling efficient scene understanding across different computing platforms. In pursuit of effective cross-modal video retrieval, Qi et al. [28] proposed a binary representation learning framework, Semantics-aware Spatial-temporal Binaries ($S^2$Bin), which captures spatial-temporal context and semantic relationships. This approach facilitates the efficient generation of binary codes for videos and texts, enhancing cross-modal retrieval performance. Li et al. [29] presented a weakly-supervised approach, Order-Constrained Representation Learning (OCRL), for predicting future actions in instructional videos. By emphasizing the sequential logic of action steps within a task, OCRL addresses the semantic level of video understanding, improving prediction accuracy across various instructional video datasets. Focusing on graphical representations of instructional videos, Schiappa et al. introduced SVGraph, a self-supervised approach that utilizes narrations for semantic interpretability [30]. By leveraging cross-modal attention, SVGraph generates unified graphical structures that encapsulate the semantics of instructional content. Furthermore, it is worth noting that SVGraph suggests techniques for creating visual summaries of instructional videos without needing annotations. They achieve this by combining visual, audio, and text cues using cross-modal attention. Additionally, an approach called $OR^2G$ [31] focuses on recognizing how object attributes change over time for better action recognition.

### B. Process mining on multimodal data

The emergence of process mining from videos is gaining attention in process analytics, aiming to extract valuable process-related insights from video data. Knoch et al. (2020) introduced an unsupervised method for process discovery from video recordings of manual assembly tasks, showcasing practical applications in industrial settings [32]. Kratsch et al. proposed the *ViProMiRA* reference architecture for leveraging video data in process mining, providing a structured approach to transforming raw video data into event logs for analysis, thus broadening the scope for exploring complex processes [33]. Lepsien et al. (2022) applied process mining to surveillance videos in pigpens, emphasizing the need for further implementation and domain-specific knowledge, presenting an abstract pipeline for process mining on video data [34]. They utilized object tracking, spatio-temporal action detection, and event abstraction techniques to translate video data into higher-level event data [35]. Furthermore, Chen et al. focused on comparing processes with Petri-net models obtained from videos [36].

Process mining extends to sensor data, exemplified by Rebmann et al.'s multi-modal approach to activity recognition and process discovery, combining motion sensor and video data for enhanced accuracy [37]. Janssen et al. introduced a method for process model discovery from smart home and IoT sensor event data, demonstrating the potential of sensor activations in mapping human routines through process mining [38].

Despite these advancements, challenges remain in aligning recognized semantics with business logic, transforming processes into business-relevant metrics, and effectively utilizing large language model capabilities in multimodal process mining, as highlighted by [39].

## VI. Conclusion

This paper presents an approach to enhancing digital traceability in manually intensive business processes with limited digital footprints or complete invisibility for an IT system. We defined concepts of Multimodal Process Mining and Multimodal Compliance/Conformance Analysis. The case study conducted as part of this research underscores the practicality and effectiveness of our approach, particularly in contexts where manual processes are predominant and difficult to trace. Our work highlights the importance of integrating advanced AI techniques with human oversight to ensure accuracy and reliability in process modeling. The scalable nature of *Vid2Log* enables both real-time and post-event process analysis. In future work, we aim to refine and expand this approach, exploring additional modalities such as depth sensing and audio. We will also explore the applicability and effectiveness of modeling mined processes across a wide range of industries and processes.

## References


[1] Q. Garrido, M. Assran, N. Ballas, A. Bardes, L. Najman, and Y. LeCun, "Learning and leveraging world models in visual representation learning," 2024.

[2] W. M. van der Aalst, A. Adriansyah, A. A. D. Medeiros *et al.*, "Process mining manifesto," in *BPM 2011: Business Process Management Workshops*, 2011, pp. 169–194.

[3] L. Zimmermann, F. Zerbato, and B. Weber, "What makes life for process mining analysts difficult? a reflection of challenges," *Software and Systems Modeling*, 2023. [Online]. Available: https://doi.org/10.1007/s10270-023-01134-0

[4] J. Michael, D. Bork, M. Wimmer, and H. C. Mayr, "Quo vadis modeling?" *Software and Systems Modeling*, 2023. [Online]. Available: https://doi.org/10.1007/s10270-023-01128-y

[5] W. M. van der Aalst, "Foundations of process discovery," in *Process Mining Handbook*. Springer, 2022, pp. 37–75.

[6] P. van Bommel, H. A. Proper, and T. P. van der Weide, "Information coverage in advisory brokers," *International Journal of Intelligent Systems*, vol. 22, no. 11, pp. 1155–1188, November 2007. [Online]. Available: http://dx.doi.org/10.1002/int.20240

[7] A. Gavric, "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] W. M. van der Aalst and S. Dustdar, "Process mining put into context," *IEEE Internet Computing*, vol. 16, pp. 82–86, 2012.

[9] W. M. van der Aalst, "Process mining: Overview and opportunities," *ACM Trans. Manag. Inf. Syst.*, vol. 3, pp. 7:1–7:17, 2012.

[10] S. Srivastava and R. Bhatnagar, "A study about process mining," *EngRN: Computer-Aided Engineering (Topic)*, 2019.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[12] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[13] IKEA, "Ikea assembly guides," https://www.ikea.com/gb/en/customer-service/product-support/assembly-guides/, 2024, accessed: 2024-06-01.

[14] Fluxicon, "Disco: Process mining software," https://fluxicon.com/disco/, accessed on June 7, 2024.

[15] OpenAI, "Dall·e 3," https://openai.com/index/dall-e-3/, 2024, accessed: 2024-06-01.

[16] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council." [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj

[17] A. Adriansyah and J. Buijs, "Mining process performance from event logs," in *International Conference on Business Process Management*. Springer, 2012, pp. 217–218.

[18] S. B. Sonawane and R. P. Patki, "Process mining by using event logs," *International Journal of Computer Applications*, vol. 116, pp. 31–35, 2015.

[19] N. Martin, L. Pufahl, and F. Mannhardt, "Detection of batch activities from event logs," *Inf. Syst.*, vol. 95, p. 101642, 2021.

[20] M. Pegoraro and W. M. van der Aalst, "Mining uncertain event data in process mining," in *2019 International Conference on Process Mining (ICPM)*, 2019, pp. 89–96.

[21] P. M. Dixit, S. Suriadi, R. Andrews, M. Wynn, A. Hofstede, J. Buijs, and W. M. van der Aalst, "Detection and interactive repair of event ordering imperfection in process logs," in *International Conference on Advanced Information Systems Engineering*. Springer, 2018, pp. 274–290.

[22] W. Song, X. Xia, H. Jacobsen, P. Zhang, and H. Hu, "Efficient alignment between event logs and process models," *IEEE Transactions on Services Computing*, vol. 10, pp. 136–149, 2017.

[23] N. I. Utama, R. A. Sutrisnowati, I. M. Kamal, H. Bae, and Y.-J. Park, "Mining shift work operation from event logs," *Applied Sciences*, 2020.

[24] H. Marin-Castro and E. Tello-Leal, "Event log preprocessing for process mining: A review," *Applied Sciences*, 2021.

[25] W. Song, X. Xia, H. Jacobsen, P. Zhang, and H. Hu, "Heuristic recovery of missing events in process logs," in *2015 IEEE International Conference on Web Services*, 2015, pp. 105–112.

[26] H. A. Proper and P. D. Bruza, "What is information discovery about?" *Journal of the American Society for Information Science*, vol. 50, no. 9, pp. 737–750, July 1999. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:9¡737::AID-ASI2¿3.0.CO;2-C

[27] F. Tosi, F. Aleotti, P. Z. Ramirez, M. Poggi, S. Salti, L. D. Stefano, and S. Mattoccia, "Distilled semantics for comprehensive scene understanding from videos," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4653–4664.

[28] M. Qi, J. Qin, Y. Yang, Y. Wang, and J. Luo, "Semantics-aware spatial-temporal binaries for cross-modal video retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 2989–3004, 2021.

[29] M. Li, L. Chen, J. Lu, J. Feng, and J. Zhou, "Order-constrained representation learning for instructional video prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, pp. 5438–5452, 2022.

[30] M. C. Schiappa and Y. Rawat, "Svgraph: Learning semantic graphs from instructional videos," in *2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM)*, 2022, pp. 45–52.

[31] Y. Ou, L. Mi, and Z. Chen, "Object-relation reasoning graph for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 133–20 142.

[32] S. Knoch, S. Ponpathirkoottam, and T. Schwartz, "Video-to-model: unsupervised trace extraction from videos for process discovery and conformance checking in manual assembly," in *BPM 2020. LNCS*, vol. 12168. Springer, Cham, 2020, pp. 291–308.

[33] W. Kratsch, F. König, and M. Röglinger, "Shedding light on blind spots - developing a reference architecture to leverage video data for process mining," *Decision Support Systems*, vol. 158, p. 113794, 2022.

[34] A. Lepsien, J. Bosselmann, A. Melfsen, and A. Koschmider, "Process mining on video data," in *ZEUS 2022, CEUR Workshop Proceedings*, vol. 3113. CEUR-WS.org, 2022, pp. 56–62. [Online]. Available: https://ceur-ws.org/Vol-3113/paper9.pdf

[35] A. Lepsien, A. Koschmider, and W. Kratsch, "Analytics pipeline for process mining on video data," in *Business Process Management Forum*, C. Di Francescomarino, A. Burattin, C. Janiesch, and S. Sadiq, Eds. Cham: Springer Nature Switzerland, 2023, pp. 196–213.

[36] S. Chen, M. Zou, R. Cao, Z. Zhao, and Q. Zeng, "Video process mining and model matching for intelligent development: Conformance checking," *Sensors*, vol. 23, no. 8, p. 3812, 2023.

[37] A. Rebmann, A. Emrich, and P. Fettke, "Enabling the discovery of manual processes using a multi-modal activity recognition approach," in *BPM 2019. LNBIP*, vol. 362. Springer, Cham, 2019, pp. 130–141.

[38] D. Janssen, F. Mannhardt, A. Koschmider, and S. van Zelst, "Process model discovery from sensor event data," in *ICPM 2020. LNBIP*, vol. 406. Springer, Cham, 2021, pp. 69–81.

[39] A. Gavric, "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*, November 28 - December 01 2023.