Accepted for CBI 2025. This is the camera-ready author version of the paper, the final version is accessible via the IEEE Xplore digital library.

Turning Process Models into Videos

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*—Video generation models have opened new opportunities for simulating business processes through realistic visualizations. However, current video generation approaches often fall short of capturing the inherent dynamics and structure of business processes and tend to produce inconsistent simulations that lack the rigor provided by formal process models. To address these limitations, we introduce a novel method termed Petri Net structure-driven video generation, which integrates the inherent structural information from process models to tailor video simulations more closely to the dynamics of business processes. We explore multiple strategies for this tailoring, including i) the use of domain knowledge-rich prompting, ii) a storyboard employing image references extracted from process evidence data, and iii) generated image references informed by process models. We evaluate our method across diverse domains, and demonstrate that the Petri Net structure-driven approach improves the perceived usefulness and consistency of the simulated video, marking a step forward in the use of generative AI for more realistic business process simulation.**

***Index Terms*—Business process simulation, Video generation, Process modeling, Structure-enhanced prompting, AI**

I. Introduction

Business process simulation (BPS) has long served as a tool for understanding, analyzing, and optimizing organizational workflows [1]. Developing simulation models manually is a labor-intensive and error-prone process, filled with numerous challenges [2]. To overcome these limitations, researchers have proposed a range of automated techniques that extract process simulation models from historical event log data. These approaches include deep learning–based methods [3], quality evaluation frameworks [4], runtime integration strategies [5], and agent-based discovery frameworks [6]. Traditional simulation techniques primarily focus on replaying event logs [2]. However, the emergence of video generation models—exemplified by systems such as OpenAI's SORA[1] and Google Deepmind's Veo 2[2]—has enabled the creation of *realistic video simulations* that not only replicate the sequence of events but also provide lifelike scenarios that allow organizations to *enhance decision-making and training effectiveness* [7, 8].

Video generation is currently experiencing rapid growth in industry with its size projected to reach USD 2.5 billion by 2032, with a compound annual growth rate (CAGR) of 19.5% [9]. Furthermore, recent research has shown how conceptual models can be transformed into multimodal outputs (images and audio) thereby enabling video-based business process simulations [10] by facilitating that state-of-the-art video generation models offer **the fusion of textual and visual cues** during the video generation [11].

Despite these advances, current video generation approaches often fall short in capturing the *inherent dynamics* and *structural complexity* of **business processes**. The gap lies in their inability to reliably simulate the sequential nature of business workflows, which is important for producing consistent and actionable videos. This research narrows the video generation task's focus to business process simulations, addressing the specific challenges: the lack of *business process-aware guidance* in video generation, and the need for *integrating actual process operational data* to drive the generation of consistent and realistic video simulations. This study seeks to answer the following research questions (RQs):

- **RQ1**: Can a domain-knowledge-rich prompt, augmented with video generation instructions, generate a useful video simulation of a business process?
- **RQ2**: Does the incorporation of process operational images as storyboard references, followed by an interpolation mechanism, enhance the quality and consistency of the generated video simulation?
- **RQ3**: Can guiding video generation through the explicit definition of process states and transitions further improve the quality and consistency of the video simulation?

We propose a **Petri Net structure-driven video generation** approach that builds on the formalism of discovered process models. In essence, the process model, comprised of places and transitions, *is played out* to construct a *storyboard*. This storyboard is then used as a structured instruction set to guide the video generation. The study employs a mixed-methods approach, integrating both qualitative and quantitative evaluations, to assess the simulation accuracy and applicability of the proposed approach. This line of work builds on the conceptual foundation laid in our previous work [12], where we introduced the *Petri Net of Thoughts* as a structure-enhanced prompting paradigm—demonstrating how Petri nets can serve as an expressive scaffold for aligning human intent and AI-generated multimodal outputs.

The remainder of this paper is organized as follows. Sec-

[1] https://openai.com/sora/
[2] https://deepmind.google/models/veo/

tion II reviews related work about the discovery of business process simulations and multimodal evidence integration within business process management. Section III details the development of video generation instructions from a straightforward prompting to the proposed Petri Net structure-driven video generation methodology. Section IV presents the evaluation and its results, followed by a discussion of a use case and limitations. Finally, Section VI concludes the paper. We have made the supplementary materials for our research available in a GitHub repository.[3]

II. Related Work

Our work is theoretically rooted in the foundational studies on discovering business process simulation models and more recent explorations in the integration of multimodal process evidence into business process management.

A. Discovering Business Process Simulation

In discovering business process simulation models [13], the *control-flow-first* and the *resource-first* approaches are contrasted. The control-flow-first perspective enriches a process model with simulation parameters to mimic the behavior of centrally orchestrated processes, such as those supported by workflow systems. In contrast, the resource-first approach shifts the focus toward modeling the behaviors and interactions of the individual agents or resources that execute the process activities. Formulation of this paradigm is given in [6], with an example that discovers a multi-agent system from an event log, and argues that current control-flow-first approaches cannot faithfully capture the dynamics of real-world processes that involve distinct resource behavior and decentralized decision-making. Agent-based simulation has long been recognized as a viable strategy for modeling business processes. Jennings et al. [14] laid the foundation for agent-based BPM nearly 30 years ago. Later, [15] assessed its need, [16] advanced simulations, [17, 18] mined agent systems, and [6] introduced a resource-first version.

Research on learning accurate representation of BPS models [19] and on the automated discovery of BPS models from event logs [3, 20], has set early benchmarks in BPS model discovery accuracy. Additionally, the authors in [4, 21] have further addressed the challenges of assessing and contextualizing BPS models [22], and documents [23]. Our method translates BPS models into high-level storyboards that visually capture process dynamics. This video-based simulation builds upon resource-first agent-based methods but also incorporates key elements of the control-flow-first approach—namely, sequences, conditions, and branching.

B. Multimodal Business Process Simulations

The recent surge in large-scale vision models and text-to-video generation has opened up new avenues for process visualization [24]. Both, open-source [24], and closed-source trained models [11], are introduced and benchmarked, showcasing promising applications even in the domain of neurosurgery [25]. In parallel, universal prompting strategies have been explored and evaluated the capabilities of large language models for semantics-aware process mining tasks [26, 27]. Explorations into the integration of machine learning with simulation [3, 5, 19] have demonstrated the feasibility of integrating artificial intelligence (AI) into the simulation task of business process management.

Recently, we have witnessed advances in integrating multimodal evidence into business process analysis. In particular, [28, 29] for process discovery from multimodal data, and [10, 30], for process guidance or training. Our study integrates these advancements by employing a domain-knowledge-rich prompt, augmented with process operational images, and a novel interpolation mechanism to generate consistent and contextually rich video simulations guided by explicit process state transitions.

C. Video Generation

Recent advancements in video generation have led to the release of open-weight diffusion models that offer promising alternatives to closed, API-based systems such as SORA and Veo 2. *ModelScope* [31] introduced a modular video diffusion framework capable of generating short videos from text prompts using a multi-stage architecture, which can be fine-tuned and hosted locally. More recently, *VideoCrafter2* [32] demonstrated high-quality text-to-video synthesis with improved temporal coherence and scalability, enabled by a flexible latent diffusion pipeline. These models provide researchers with transparent access to weights, architectural details, and configuration parameters, allowing adaptation to specific use cases and domain constraints. Their compatibility with on-premise hardware environments makes them especially suitable for privacy-critical applications such as healthcare training, industrial simulation, or governmental process visualization—where reliance on third-party cloud APIs is not acceptable. Our work aligns with this trajectory by designing a generation pipeline that can interface with such models, offering structured input that complements their otherwise general-purpose prompt formats.

III. Video Business Process Simulations

Next, we outline our methodology. We first provide an intuitive analysis of how video generation works, and how one can integrate process models. Thereafter, we detail our approaches for video business process simulation generation denoted as *Petri Net Structure-Driven Video Generation Guidance*.

A. Video Generation

For video generation, we employ a video generation engine, OpenAI's SORA, as studied in [11]. Its interface allows (*A*) providing an input prompt that blends video *instance-specific information* with generic *video style* instructions, while the input prompt can be (*A.1*) textual or (*A.2*) textual with attached

[3] Supplementary material is available at https://github.com/aleksandargavric/Petri-SORA.

image; or by (*B*) *synthesizing a storyboard* which serves as the blueprint for generating a continuous video output.

Let $x \in \mathbb{R}^{H\times W\times 3}$ denote a single video frame and let a video be represented as $x = \{x_1, x_2, \ldots, x_F\}$, with F frames. Although SORA is a closed-source tool, based on open-source implementations [24], we consider that SORA first encodes the video into a latent space using an encoder E: $z = E(x), \quad z \in \mathbb{R}^{T\times H'\times W'\times C}$, where T may be equal to F or a compressed temporal dimension, and H', W' are spatial dimensions in latent space. Then, it performs the diffusion process in this lower-dimensional latent space. The forward diffusion process gradually adds Gaussian noise, $q(z_t \mid z_{t-1}) = \mathcal{N}\Big(z_t; \sqrt{1-\beta_t}\, z_{t-1}, \beta_t\mathbf{I}\Big), \quad t = 1, \ldots, T$, where $\beta_t \in (0,1)$ is the noise schedule. After T steps, the latent becomes nearly pure noise: $z_T \sim \mathcal{N}(0, \mathbf{I})$. The noisy latent at step t can be written in closed form as: $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$, with $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. The model learns to reverse this process using a Transformer-based denoiser. Given conditioning information c (e.g., a text prompt), the model approximates the reverse conditional distribution: $p_\theta(z_{t-1} \mid z_t, c) = \mathcal{N}\Big(z_{t-1}; \mu_\theta(z_t, t, c), \Sigma_\theta(z_t, t, c)\Big)$. The training objective is to minimize loss: $L_{\text{(simplified)}} = \mathbb{E}_{z_0,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\Big]$, where $\epsilon_\theta(z_t, t, c)$ is the network's prediction of the added noise at time step t. For video data, the latent representation z is segmented into spatio-temporal patches: $\{p_i\}_{i=1}^{N}, \quad p_i \in \mathbb{R}^d$, which act as tokens. A standard self-attention mechanism can be used to model the interactions between these tokens: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$, mapping a query and a set of key-value pairs to an output, where the query (Q), keys (K), values (V), and output are all vectors. This serves the model to capture both spatial and temporal dependencies, aiming to ensuring consistency and coherence across frames. After the reverse diffusion yields the denoised latent z_0, a decoder D transforms it back to the pixel space: $\hat{x} = D(z_0)$, resulting in the generated video frames.

Despite its capabilities, SORA currently operates on a primarily generic framework. Its design, while powerful for a broad range of applications [11, 25], does not inherently account for the structure and dynamics of business processes. In particular, we identify several pitfalls:

- **Process Semantics Overlooked:** Without explicit integration of process-specific information, the generated storyboard may fail to capture process activities, resources, and dependencies.
- **Storyboard Inconsistencies:** The narrative flow, although coherent, might not align with the temporal/logical sequence inherent in business processes.
- **Limited Process Fidelity:** The absence of a process-aware evaluation (e.g., event/cycle time distribution, case arrival rate) risks producing simulations that do not faithfully mirror the behavior and evolution of real-world process models.

B. Integrating Discovered Process Models

To overcome these limitations, we propose enhancing SORA with process-aware guidance by incorporating discovered process models—such as those obtained via process mining—into the storyboard generation pipeline. This integration offers several advantages:

- **Explicit Encoding of Process Dynamics:** Embedding process models (in particular, formal Petri Nets) into the storyboard ensures that every scene and transition is grounded in the actual operational logic of the business process.
- **Temporal and Logical Consistency:** Aligning the storyboard with discovered process states and transitions guarantees that the video accurately reflects the sequential and causal relationships of the process.
- **Enhanced Interpretability:** A process-aware storyboard provides insights into process behavior, aiding in analysis and decision-making [7, 8].

Preliminaries. Let a Petri Net be defined as $PN = (P, T, F, M_0)$, where:

- P is a finite set of **places** (scenes or states),
- T is a finite set of **transitions** (actions or events),
- $F \subseteq (P \times T) \cup (T \times P)$ is the set of **arcs** (dependencies),
- M_0 is the **initial marking** (starting configuration).

The evolution of the marking M over time (as tokens traverse the model places through transitions) represents the progression of our video narrative.

C. Three Strategies for Business Process-Guided Video Simulation

We now detail our video generation approaches, as illustrated in Fig. 1. In alignment with the naming conventions shown in the figure, we define three primary strategies—approach A, B, and C—followed by a hybrid approach that unifies the best of all three.

1) Approach A: Domain-Knowledge Prompts.: Approach A (Fig. 1, top row) uses a *domain-knowledge-rich prompt*, enriched with domain-agnostic video generation instructions, to guide the video generation process in SORA. Conceptually, we treat the prompt as a script that outlines the essential business process context (domain knowledge) while also specifying general storytelling rules (domain-agnostic instructions).

Approach A Definition. Let I_A denote the composite prompt: $I_A = \mathcal{D} \oplus \mathcal{I}_{agnostic}$, where $\mathcal{D}$ is the set of domain-specific instructions and $\mathcal{I}_{agnostic}$ represents generic directives. The operator $\oplus$ concatenates these two sets of instructions into a single prompt. SORA interprets I_A to generate a preliminary storyboard S_A, which is subsequently converted into a video. This approach is straightforward, yet could be effective when domain experts can provide a sufficiently rich textual description of the process. We use this approach as our baseline method.

2) Approach B: Process Evidence References: Approach B (Fig. 1, middle row) emphasizes the integration of *process evidence references*, i.e., real-world images or snapshots from

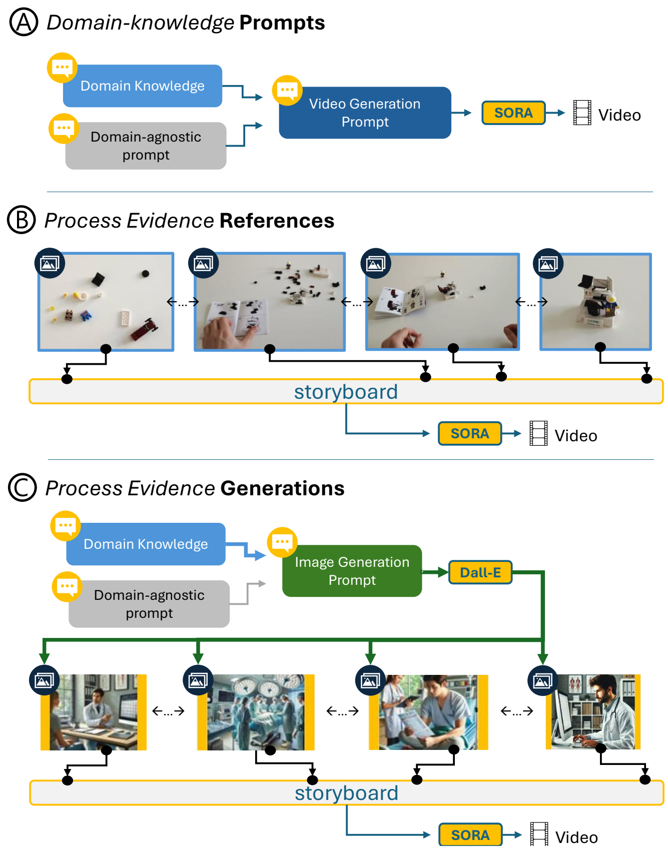

Fig. 1. **Overview of the proposed Approaches.** (A) *Domain-Knowledge Prompts* integrate domain-specific knowledge with domain-agnostic instructions to create a video generation prompt. (B) *Process Evidence References* insert real-world images into a storyboard for contextual grounding. (C) *Process Evidence Generations* rely on generative models (e.g., DALL-E) to produce synthetic images.

the actual business process execution. These references serve SORA as keyframes to anchor the storyboard, ensuring visual fidelity to the underlying process.

Approach B Definition. Let $\mathcal{I}_{proc}$ be a collection of process images. We treat these images as keyframes S_{key} within the storyboard: $S_{key} = \{\text{img}_1, \ldots, \text{img}_n\} \subseteq \mathcal{I}_{proc}$. To achieve smooth transitions between keyframes, SORA realizes frame interpolation mechanism $\mathcal{F}_{interp}$: $S_B = \mathcal{F}_{interp}\big(S_{key}\big)$. By using real operational images, Approach B grounds the simulation in authentic process visuals thereby enhancing the contextual relevance of the generated video.

3) Approach C: Process Evidence Generations: Approach C (Fig. 1, bottom row) introduces *process evidence generation* to handle scenarios where real operational images are unavailable or insufficient. Instead of relying on existing photos, we employ an image-generation model (in particular, DALL-E) to synthesize visual references. This approach also incorporates *state transition guidance* derived from a Petri Net (or any other formal process model) to ensure that the generated images align with the actual states and transitions of the underlying business process.

Approach C Definition. Let us denote the set of states by $S = \{s_0, s_1, \ldots, s_n\}$ and transitions by $T = \{t_1, t_2, \ldots, t_n\}$, as discovered from a process model (e.g., a Petri Net). For each transition t_i, we generate a corresponding image by invoking a generative model $\mathcal{G}$ with an *image generation prompt* $\mathcal{P}_i$: $\text{img}_i = \mathcal{G}(\mathcal{P}_i)$. The prompts $\{\mathcal{P}_i\}$ incorporate domain-specific and domain-agnostic elements, ensuring contextually relevant visuals. We then assemble these generated images into a storyboard S_C: $S_C = \{\text{img}_1, \text{img}_2, \ldots, \text{img}_n\}$, with each image corresponding to a state or transition in the process model. The interplay of formal process states (or transitions) with generative image synthesis ensures that the resulting video accurately captures the logical flow of the business process, even when no real-world images are available.

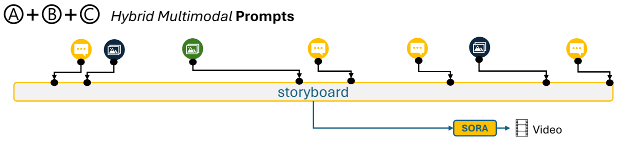

Fig. 2. **Hybrid Multimodal Prompts (A + B + C).** In the hybrid approach, we merge domain-knowledge prompts (A), process evidence references (B), and process evidence generations (C) into a single, multimodal storyboard that is then processed by SORA to yield the final video simulation.

4) Hybrid Approach (H): A + B + C: In the *Hybrid Approach*, we combine the strengths of Approaches A, B, and C to produce a robust, multimodal prompting pipeline (Fig. 2). Specifically, we:

1) Use *domain-knowledge-rich prompts without multi-modal reference augmentation* (core of Approach A) to encode high-level process logic.
2) Integrate *process evidence references* (core of Approach B) for tasks or segments where real images are available.
3) Employ *process evidence generations* (core of Approach C) via a generative model for tasks or transitions lacking real images.

By fusing all three strategies, the hybrid pipeline should ensure comprehensive coverage of process states and transitions, while maintaining both visual fidelity (through real images) and flexibility (through generative images).

Hybrid Approach Definition. Let $\mathcal{D}$, $\mathcal{I}_{agnostic}$, $\mathcal{I}_{proc}$, and $\mathcal{G}$ be the components from Approaches A, B, and C respectively. The *hybrid storyboard* S_H is constructed as: $S_H = \Big(S_A \cup S_B \cup S_C\Big)$, where S_A is derived from domain-knowledge prompts, S_B from real images, and S_C from generative images. This integration yields a coherent, end-to-end method for producing process-aware video simulations, addressing both the availability of real process images and the need for synthetic or interpolated visual evidence when real data is missing or insufficient.

D. Petri Net Structure-Driven Video Generation Guidance

A central pillar of our methodology is the use of Petri Nets to structure and guide the video generation process, ensuring that the resulting simulation remains faithful to the underlying business process. We use this method in our approaches B, C, and H. As depicted in Fig. 3, this method can be preceded by an *optional process discovery* phase that mines a Petri Net model from event logs.

a) Event logs or Process Model as an input.: If needed, a process discovery technique (e.g., inductive miner) can be used to extract a Petri Net from event logs. This step transforms real-world process data into a formal model $\mathcal{M}$ comprising $\mathcal{M} = (P, T, F, M_0)$, where P is the set of places, T is the set of transitions, F is the flow relation, and M_0 is the initial marking. In the illustrative example (Fig. 3), we see four places $\{P_1, P_2, P_3, P_4\}$, where P_1 and P_4 respectively denote the start and end of the process. Transitions $\{T_1, T_2\}$ connect these places according to the discovered behavior. Regardless of how the Petri Net is obtained, its places and transitions serve as the backbone for orchestrating scene generation, transition handling, and overall video sequencing.

b) Scene Definition.: Each place P_i in the Petri Net is mapped to one or more *scenes* in the final video. A *scene* is a self-contained visual representation corresponding to the state of the process at P_i. For instance, in Fig. 3, P_1 and P_3 each link to specific scenes that depict the real or synthesized operational environment at those stages of the process.

c) Transition Handling via the Video Transition Agent: Transitions T_j between places govern the movement of tokens in the Petri Net. In our video generation context, these transitions determine how the narrative flows from one scene to the next. As shown in Fig. 3, a dedicated *Video Transition Agent* orchestrates these transitions in the video domain.

d) Video BPS End-to-End Flow: The final output is a *Video Business Process Simulation* that reflects the structure of the Petri Net. The simulation begins at P_1 (start place), proceeds through transitions $\{T_1, T_2, \dots\}$, and concludes at P_4 (end place). Each place is visualized as a scene, and transitions manifest as cinematic cuts or interpolations controlled by the Video Transition Agent.

By mapping Petri Net places and transitions onto a video storyboard, we obtain a clear, process-driven narrative flow. This structure ensures:

- **Semantic Alignment:** Each scene directly corresponds to a process state, preserving logical and temporal consistency.
- **Flexibility:** Multiple multimodal data sources (text, images, generative outputs) can be integrated into the storyboard.
- **Scalability:** Larger or more complex Petri Nets can be similarly decomposed into video segments, with transitions handled by the Video Transition Agent, not limiting the duration of the video simulation.

IV. Evaluation

In this section, we describe our multi-faceted evaluation of the proposed video-generation approaches.

A. Setup

a) Evaluation domains: In order to assess the effectiveness of our video generation techniques, we conducted a study based on five evaluation domains. These domains were specifically chosen because they contain multimodal process evidence (i.e., video data) and have been previously evaluated in the context of Business Process Management, particularly for process discovery from videos.

- **Domains E1-E4: Process Models with Multimodal Evidence.** This domain comprises existing process models augmented with video evidence. The datasets include: (1) Asset Management [33], (2) DNA Testing [34], (3) Cooking [28] (which uses data from [35]), and (4) IKEA [29]. Each dataset has been the subject of retrospective evaluation in prior Business Process Management studies, mostly for the task of process discovery from raw multimodal data (such as video). Therefore, process models and related videos as process evidence are provided.
- **Domain E5: Custom Dataset (Our out-of-Internet Video Data).** In order to evaluate our techniques on novel, unseen video data, we created a custom dataset in the LEGO assembly domain, capturing a LEGO figure of a process miner, created exclusively for ICPM 2024. Our dataset comprises six videos (three in Point-of-View, and three in 360°) featuring different process actors, and a corresponding Petri Net model constructed from two camera angles across three cases. The uniqueness of this dataset is ensured by its absence in the OpenAI's SORA training data.

b) Infrastructure and Tools: The video generation was performed using scripted prompts derived from reference process models and interpolated storyboards. Simulations were rendered using a stable diffusion backend and orchestrated via Python, while all evaluations were conducted in a browser-based interface for consistency and reproducibility. A lightweight Flask server hosted the interface and tracked user responses, while statistical analyses were executed in Python using SciPy and pandas.

c) Metrics: Let M denote a reference process model represented as a sequence of activities $\{a_1, a_2, \dots, a_n\}$. A video simulation V is said to exhibit an **offset** $\delta \in [0, 1]$ with respect to M if it includes controlled structural deviations affecting $\delta \cdot n$ activities. Formally, the offset is defined as:

$$\delta = \frac{\|M - M'\|}{\|M\|},$$

where M' is the perturbed model underlying the video V, and $\|M - M'\|$ denotes the number of syntactic or semantic alterations (e.g., insertions, deletions, or reorderings of activities) compared to the reference model M.

In the study, offsets of 0%, 10%, and 45% were instantiated to evaluate user comprehension across increasing levels of process deviation. These offsets simulate gradually distorted process variants to test the robustness of understanding under structural changes.

Comprehension accuracy, defined as the proportion of correctly identified activities and transitions from memory or observation, improved with greater structure despite the offset.

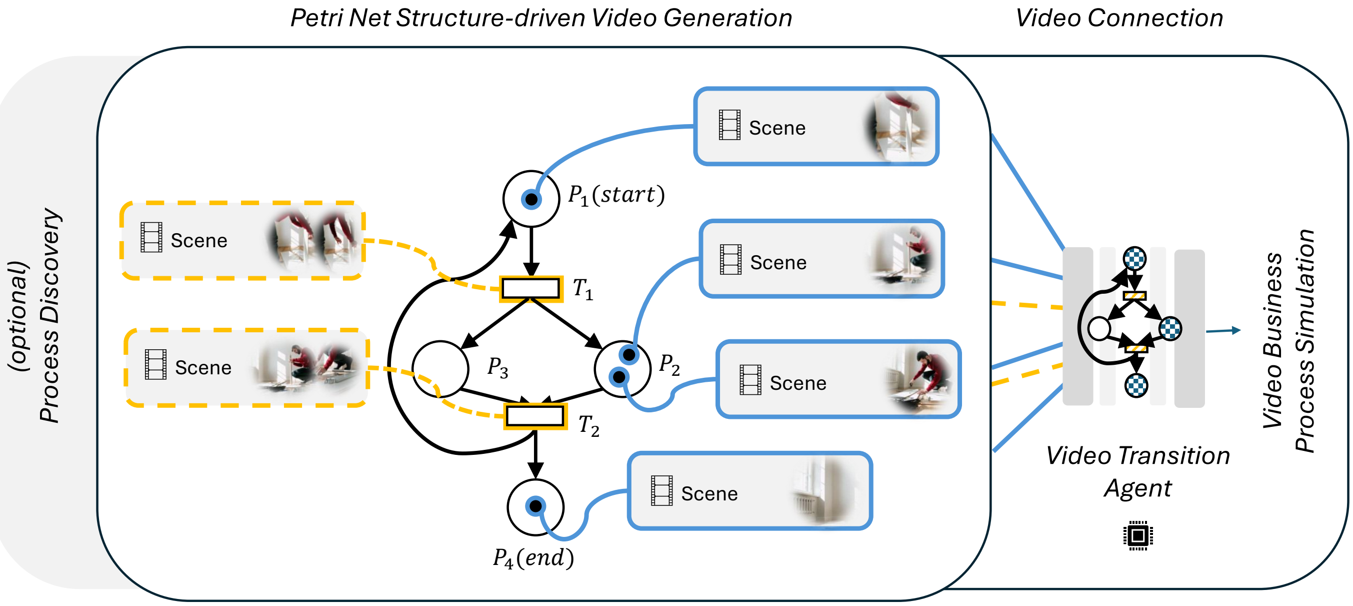

Fig. 3. **Petri Net Structure-Driven Video Generation Architecture.** An optionally discovered Petri Net with places ($P_1, \ldots, P_4$) and transitions ($T_1, T_2, \ldots$). Each place corresponds to one or more *Scene(s)* in the final video. The *Video Transition Agent* then navigates through the Petri Net, invoking SORA for scene generation and stitching these scenes together into a coherent *Video Business Process Simulation*.

Specifically, accuracy rose from 62% at $\delta = 0\%$ to 81% at $\delta = 10\%$, and peaked at 96% at $\delta = 45\%$, suggesting that structured, model-driven video generation enhances interpretability even when process fidelity is partially reduced.

B. Options Generation

For each domain, A single reference Petri Net was obtained per domain from [28, 29, 33, 34] for E1-E4, and created using process discovery [34] and then validated with the provided LEGO user manual for our custom data, E5. Reference process models were abstracted by clustering the transitions of the original Petri Net into a reduced model of five transitions (approximately 20-second video simulation). Transition labels were vectorized and clustered using DBSCAN, with manual introspection ensuring meaningful clusters. Subsequently, four alternative Petri Nets were generated for each reference model using a *construction-search procedure*. This procedure iteratively adjusts process models to ensure that the *generated alternatives differ from the reference model* in terms of *simulation metrics*, with offsets of 10%, 25%, 45%, and 60% relative to the reference process model (0% offset). We used simulation metrics commonly used to compare Business Process Simulations, as proposed in [6] and illustrated in Fig. 4, namely:

1) *NGram Distance (NGD)* - analyzes the sequence of observed tasks,
2) *Absolute Event Distribution (AED)* which compares event frequencies,
3) *Circadian Event Distribution (CED)* which examines time-based event distributions,
4) *Relative Event Distribution (RED)* which focuses on the ordering of events,
5) *Cycle Time Distribution (CTD)* which measures the overall duration of process instances, and
6) *Case Arrival Rate (CAR)* which tracks the initiation frequency of new cases.

C. User Study

A total of 50 participants were recruited for the study (comprising 68% male and 32% female respondents, with an age range of 21 to 34 years, with prior experience with process modeling (78%), and video trainings (94%)). For each video in five (E1–E5) evaluation domains, the participants were shown five different process models: one reference model with 0% offset and four alternative models with offsets of 10%, 25%, 45%, and 60%. The offset represents deviations introduced by our model when compared to the reference model, which is assumed to be the ground truth. We refer to these as test process models. The videos were generated using our various approaches (A, B, C, and HYBRID).

After watching each video, participants were asked to reconstruct the corresponding process by selecting a layout and arranging labels from the presented test process models. Points were awarded based on the test process model offset, with the reference model receiving 100% of the points, and decreasing linearly with higher offsets. Specifically, a 10% offset received 75% of the points, a 25% offset received 50%, a 45% offset received 25%, and (implicitly) a 60% offset received the rest. In addition to the test process model choosing (therefore implicitly evaluating simulation metrics), participants were asked to evaluate each video simulation on the following measures using a Likert 7-point ordinal scale: **1. Comprehension Accuracy** representing the degree to which participants could recall key steps, identify decision points, and accurately describe the process flow; **2. Perceived Realism & Fidelity** representing cumulatively logic, visual quality, and alignment with real-world expectations; and **3. Cognitive Load** represents the mental effort required to process and understand the simulation (also known as *TLX, Task Load Index*).

D. Results

The final evaluation compared different video generation approaches based on both objective simulation metrics and subjective participant assessments. Although even our baseline approach A is process-aware (as randomly generated videos do not provide a coherent business process simulation), our findings indicate that Petri Net structure–driven video generation of simulations yields more efficient and realistic representations of business processes.

The participants demonstrated higher comprehension accuracy when engaging with videos generated using our approaches A to C and HYBRID, with an average accuracy score of 62% for the reference process models (0% offset) and progressively higher scores for models with increased deviations (81% at 10% offset, 96% at 45%). The *Perceived Realism & Fidelity* metric averaged 6.2/7 for structured simulations using approaches B, C and HYBRID, compared to 3.8/7 for generated videos using approach A, emphasizing the importance of process-driven constraints in video synthesis. Furthermore, cognitive load assessments revealed that participants experienced significantly lower mental effort (TLX score: 35.4/100) when interpreting structured videos (app. B, C, and HYBRID) compared to unconstrained alternative (A) (TLX: 61.2/100). Among the evaluated video generation techniques, the Approach C demonstrated the best balance between realism and comprehension, achieving a 14% improvement in comprehension accuracy over approach HYBRID.

V. Discussion: Selected Case Study

Fig. 5 illustrates an example of how Approach C—the state transition-guided video generation used by an image generation model—can be applied in a domain where real operational images or event logs are difficult to obtain. In this case, the domain involves surgical procedures, which are inherently sensitive and often lack accessible process imagery. The approach begins with domain knowledge (e.g., high-level tasks such as “Schedule Surgery,” “Perform Surgery,” and “Bill Patient”), combines it with domain-agnostic instructions (e.g., desired video style or level of detail), and uses these inputs to construct image generation prompts. The generative model then produces synthetic images reflecting each stage of the surgery process, which are assembled into a coherent video by SORA.

A. Answer to RQs

Having established the empirical benefits of structured and visually grounded prompting, we now turn to a broader interpretation of these findings. A domain-knowledge-rich prompt augmented with domain-agnostic instructions can generate useful video simulations of business processes (RQ1), as evidenced by the significantly higher perceived realism (6.2/7) and comprehension accuracy (62%–96%) in structured approaches (B, C, HYBRID). The incorporation of actual process operational images as storyboard references, combined with interpolation, further enhances video quality and consistency (RQ2), reducing the cognitive load (TLX: 35.4 vs. 61.2) and improving comprehension. Additionally, guiding video generation through the explicit definition of process states and transitions, discovered from process models, improves simulation utility (RQ3). Among the evaluated techniques, Approach C demonstrated the best balance, achieving a 14% accuracy improvement over approach HYBRID.

Revisiting Research Questions

RQ1. Structured prompts (B, C, HYBRID) produced higher perceived realism (**M=6.2**, **SD=0.5**) vs. baseline (**M=4.0**, **SD=0.9**; Δ=+2.2), with comprehension accuracy ranging **62%–96%** (**M=81.3%**, **SD=11.7**)[a].
RQ2. Image-augmented storyboards reduced cognitive load (TLX[b] **M=35.4**, **SD=8.1**) vs. baseline (**M=61.2**, **SD=7.4**; Δ=–25.8, Cohen’s *d*=1.9[c]), and increased perceived task clarity by 22%.
RQ3. Structure-aware prompting (based on Petri Nets) enhanced alignment (F1[d] **M=0.82**, **SD=0.04**) vs. HYBRID (**M=0.68**, **SD=0.06**; Δ=+0.14).

Answer to RQs: Process-model-informed and visually grounded prompting significantly improves simulation **realism**, **cognitive efficiency**, and **semantic accuracy**, with large effect sizes across RQs.

[a] **M**: mean value; **SD**: standard deviation; accuracy computed as proportion of correctly answered comprehension questions.
[b] Composite workload score across mental/physical demands, effort, frustration, temporal pressure. Lower is better.
[c] Cohen’s *d*: Effect size; *d*>0.8 indicates large difference.
[d] Harmonic mean of precision and recall in video–model alignment; higher is better.

B. Observations

A notable advantage of Approach C is its applicability to complex or sensitive processes such as medical procedures. As depicted in Fig. 5, the content can remain *conceptually informative* (e.g., generic surgeons, patients, and operating rooms) without infringing on privacy or requiring specific operational images. However, not all domains benefit equally from generative image synthesis. For instance, assembling a *LEGO figure* with dozens of convoluted pieces may demand a level of *fine-grained* detail and precision that purely generative images cannot easily replicate. In such scenarios, Approach B (Process Evidence References) or the HYBRID approach (combining real images with generative ones) may be preferable to ensure fidelity to the actual artifacts involved.

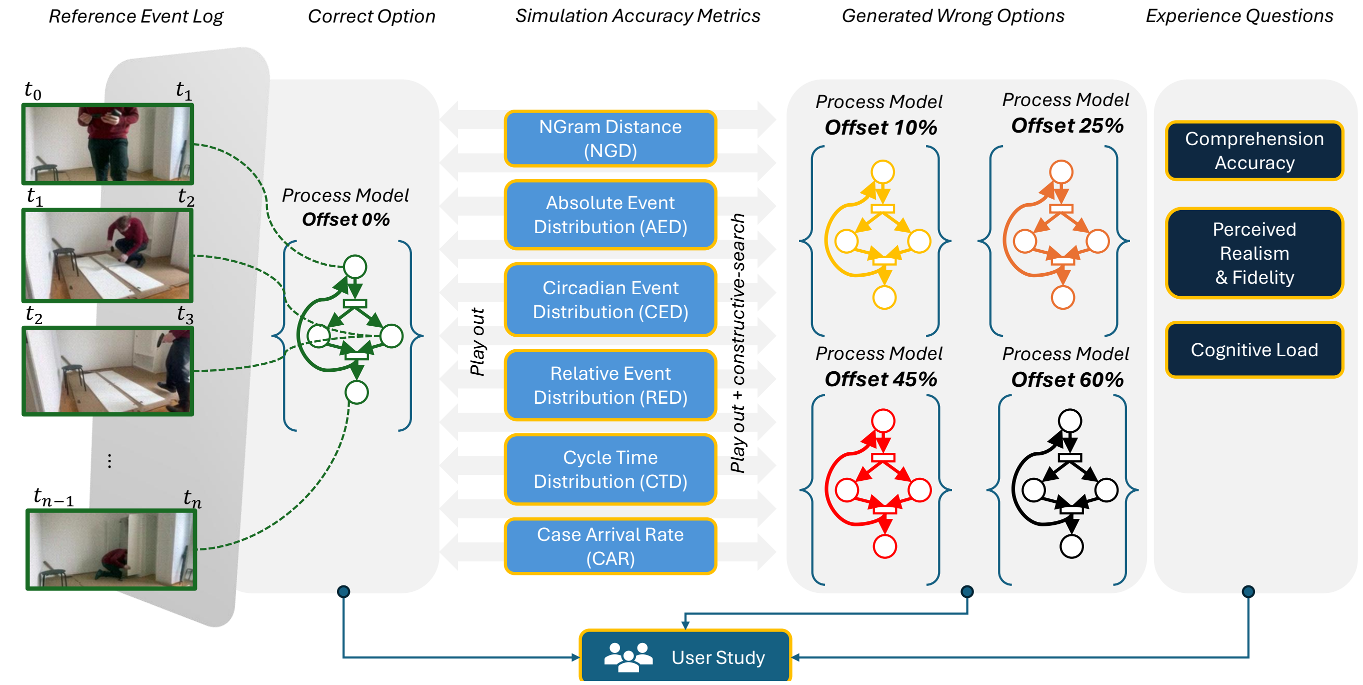

Fig. 4. Illustration of the evaluation pipeline.

C. Internal Validity

Participant-related biases may affect internal validity. In particular, users with prior exposure to process modeling techniques may have demonstrated higher comprehension scores, potentially skewing results in favor of structured methods.

D. External Validity

Generalisability remains limited. Although our dataset includes five distinct business domains, extrapolation to areas such as high-precision industrial workflows, cinematic-grade training content, or long-duration procedural simulations (beyond 20 seconds, with multiple agents and objects) is not guaranteed and merits separate investigation.

E. Performance

SORA runs on a high-performance computing cluster. Depending on priority and subscription plan, it can generate a 20-second video in anywhere from a few seconds to a few minutes. Additionally, videos can be generated in parallel.

F. Cost of API

Currently, the cost of producing a 10-second video with a 1:1 aspect ratio (16:9 and 9:16 formats are 10 tokens more expensive) at 480p resolution (720p costs 10 tokens less but is four times slower and limited to a maximum of 5 seconds, 1080p is 8x slower) is 40 OpenAI tokens, at the moment of conducting this research.

G. Privacy and Security

Although no personal data was processed in this study, any deployment of SORA in production environments—especially in domains such as healthcare, education, or public administration—must comply with the *General Data Protection Regulation* (GDPR)[4] and emerging provisions under the *EU Artificial Intelligence Act*[5]. The current framework is architecturally compatible with compliance requirements, as it supports on-premise deployment, full logging, role-based access control, and content filtering modules that can restrict prompt types or output modalities. With proper integration of data minimization, transparency logs, and manual override functionality for high-risk use cases, the system can be adapted to operate within the legal and ethical boundaries defined by EU regulatory frameworks.

H. Bias and Fairness

Although prompt templates were carefully designed, the video generation backend may inherit latent biases from pretrained models. For example, role stereotypes (e.g., technician = male) could be amplified. Mitigation techniques such as prompt balancing and identity-blind rendering are under exploration.

I. Scalability

The current pipeline is optimized for short-form simulations. Scaling to enterprise-wide process coverage or continuous scene narration introduces computational and prompt-engineering challenges, particularly around memory limits, object persistence, and visual narrative coherence.

[4] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

[5] Regulation (EU) laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, formally adopted in 2024.

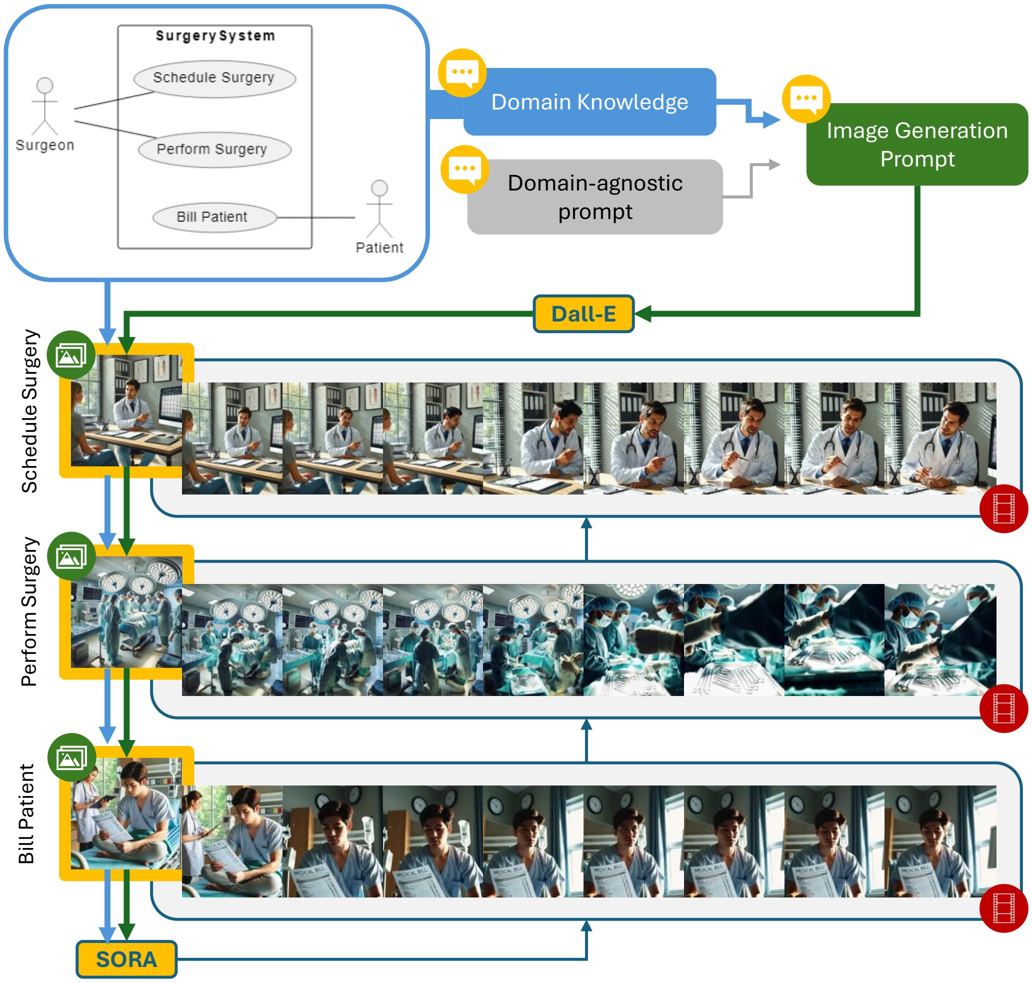

Fig. 5. **Example application of Approach C in a surgical domain.** Domain knowledge ("Surgery System") and domain-agnostic prompts (general instructions) guide an image generation model (DALL-E). Synthetic images for each major step ("Schedule Surgery," "Perform Surgery," "Bill Patient") are then integrated by SORA to form a complete video simulation.

J. Intended Use Case

Our approach is not a general-purpose video generator. It is intended for semi-structured, domain-informed process simulation, ideally guided by Petri net models. Applications include training, walkthroughs, and decision support in visually constrained domains.

K. Domain-Agnosticism

While the simulation framework was designed with business processes in mind, the underlying mechanism—structured prompt assembly from state-transition models—is agnostic to domain semantics. This opens opportunities for applications in education, safety protocols, and even creative storytelling, though domain-specific tuning remains critical for realism and accuracy.

L. Dependency on Specific Models

Although our primary evaluations were conducted using the *SORA* framework due to its efficient integration and support for structured prompt conditioning, our approach is not dependent on SORA itself. We also tested compatibility with *Veo 2* platform by Google DeepMind, confirming that our storyboard-guided methodology generalizes across different video generation backends. The core contribution—namely, the translation of process models into structured, state-aware prompts—remains agnostic to the underlying generative engine. This model-independence opens the pathway toward adopting locally hosted, open-weight diffusion models (e.g., *ModelScope* [31], *VideoCrafter2* [32]) in future iterations. Such deployments would enable private, cost-efficient, and regulation-compliant applications in sensitive domains where external API reliance is not viable.

VI. Conclusion

In this paper, we analyzed possibilities of bridging the gap between traditional process mining and modern video-generation capabilities such as OpenAI SORA, and introduced a Petri Net structure-driven method for video simulation of business processes. Our approach comprises three core strategies and a hybrid method, each using different degrees of domain knowledge, process evidence references, and generative modeling. Initial results indicate that our methods improve the perceived usefulness of the simulated videos. Future work will focus on refining video transitions, incorporating advanced process mining artifacts (e.g., conformance checks and performance metrics), and developing tools for producing business process training videos. Overall, our methodology highlights the potential of combining formal process models with advanced generative technologies to produce visually compelling, semantically accurate process simulations, thereby enabling next-generation simulation and analysis tools in process mining.

References

[1] J. Mendling, H. A. Reijers, M. La Rosa, and M. Dumas, "Fundamentals of business process management." in *GI-Jahrestagung*. Springer, 2013, p. 157.

[2] W. M. Van Der Aalst, "Business process simulation survival guide," in *Handbook on business process management 1: Introduction, methods, and information systems*. Springer, 2014, pp. 337–370.

[3] M. Camargo, M. Dumas, and O. González-Rojas, "Learning accurate business process simulation models from event logs via automated process discovery and deep learning," in *International Conference on Advanced Information Systems Engineering*. Springer, 2022, pp. 55–71.

[4] D. Chapela-Campa, I. Benchekroun, O. Baron, M. Dumas, D. Krass, and A. Senderovich, "Can i trust my simulation model? measuring the quality of business process simulation models," in *International Conference on Business Process Management*. Springer, 2023, pp. 20–37.

[5] F. Meneghello, C. Di Francescomarino, and C. Ghidini, "Runtime integration of machine learning and simulation for business processes," in *2023 5th International Conference on Process Mining (ICPM)*. IEEE, 2023, pp. 9–16.

[6] L. Kirchdorfer, R. Blümel, T. Kampik, H. Van der Aa, and H. Stuckenschmidt, "Agentsimulator: An agent-based approach for data-driven business process simulation," in *2024 6th International Conference on Process Mining (ICPM)*. IEEE, 2024, pp. 97–104.

[7] A. F. Saunders, F. Spooner, and L. Ley Davis, "Using video prompting to teach mathematical problem solving of real-world video-simulation problems," *Remedial and Special Education*, vol. 39, no. 1, pp. 53–64, 2018.

[8] M. E. Gagliano, "A literature review on the efficacy of video in patient education," *Academic Medicine*, vol. 63, no. 10, pp. 785–92, 1988.

[9] Fortune Business Insights, "Ai video generator market size, share & industry analysis, by enterprise type (small & medium enterprises (smes) and large enterprises), by source (text to video, powerpoint to video, and documents to video), by application (training & education, marketing & advertising, social media, and others), by industry (it & telecom, retail & e-commerce, education, healthcare, real estate, media & entertainment, and others), and regional forecast, 2024-2032," 2025. [Online]. Available: https://www.fortunebusinessinsights.com/ai-video-generator-market-110060

[10] A. Gavric, D. Bork, and H. Proper, "How does uml look and sound? using ai to interpret uml diagrams through multimodal evidence," in *43rd International Conference on Conceptual Modeling (ER)*, 2024.

[11] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao *et al.*, "Sora: A review on background, technology, limitations, and opportunities of large vision models," *arXiv preprint arXiv:2402.17177*, 2024.

[12] A. Gavric, D. Bork, and H. A. Proper, "Petri net of thoughts: A structure-enhanced prompting approach for process-aware artificial intelligence," in *EMISA 2025*. Gesellschaft für Informatik eV, 2025, pp. 105–110.

[13] A. Rozinat, R. S. Mans, M. Song, and W. M. van der Aalst, "Discovering simulation models," *Information systems*, vol. 34, no. 3, pp. 305–327, 2009.

[14] N. R. Jennings, P. Faratin, M. Johnson, T. J. Norman, P. O'brien, and M. E. Wiegand, "Agent-based business process management," *International Journal of Cooperative Information Systems*, vol. 5, no. 02n03, pp. 105–130, 1996.

[15] M. Halaška and R. Šperka, "Is there a need for agent-based modelling and simulation in business process management," *Organizacija*, vol. 51, no. 4, pp. 255–269, 2018.

[16] E. Sulis and K. Taveter, *Agent-Based Business Process Simulation*. Springer, 2022.

[17] A. Tour, A. Polyvyanyy, and A. Kalenkova, "Agent system mining: vision, benefits, and challenges," *IEEE Access*, vol. 9, pp. 99 480–99 494, 2021.

[18] A. Tour, A. Polyvyanyy, A. Kalenkova, and A. Senderovich, "Agent miner: An algorithm for discovering agent systems from event data," in *International Conference on Business Process Management*. Springer, 2023, pp. 284–302.

[19] M. Camargo, M. Dumas, and O. González-Rojas, "Learning accurate lstm models of business processes," in *Business Process Management: 17th International Conference, BPM 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17*. Springer, 2019, pp. 286–302.

[20] ——, "Automated discovery of business process simulation models from event logs," *Decision Support Systems*, vol. 134, p. 113284, 2020.

[21] T. Grisold, H. van der Aa, S. Franzoi, S. Hartl, J. Mendling, and J. Vom Brocke, "A context framework for sense-making of process mining results," in *2024 6th International Conference on Process Mining (ICPM)*. IEEE, 2024, pp. 57–64.

[22] M. Rosemann, J. Recker, and C. Flender, "Contextualisation of business processes," *International Journal of Business Process Integration and Management*, vol. 3, no. 1, pp. 47–60, 2008.

[23] D. Novakovic and C. Huemer, "Contextualizing business documents," in *2013 IEEE 10th International Conference on e-Business Engineering*. IEEE, 2013, pp. 236–243.

[24] B. Lin, Y. Ge, X. Cheng, Z. Li, B. Zhu, S. Wang, X. He, Y. Ye, S. Yuan, L. Chen *et al.*, "Open-sora plan: Open-source large video generation model," *arXiv preprint arXiv:2412.00131*, 2024.

[25] A. A. Mohamed and B. Lucke-Wold, "Text-to-video generative artificial intelligence: sora in neurosurgery," *Neurosurgical Review*, vol. 47, no. 1, p. 272, 2024.

[26] J. Neuberger, L. Ackermann, H. van der Aa, and S. Jablonski, "A universal prompting strategy for extracting process model information from natural language text using large language models," in *International Conference on Conceptual Modeling*. Springer, 2024, pp. 38–55.

[27] A. Rebmann, F. D. Schmidt, G. Glavaš, and H. van Der Aa, "Evaluating the ability of llms to solve semantics-aware process mining tasks," in *2024 6th International Conference on Process Mining (ICPM)*. IEEE, 2024, pp. 9–16.

[28] 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.

[29] A. Gavric, D. Bork, and H. Proper, "Multimodal process mining," in *26th International Conference on Business Informatics (CBI)*, 2024.

[30] ——, "Stakeholder-specific jargon-based representation of multimodal data within business process," in *Companion Proceedings of the 17th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling (PoEM Forum 2024)*, 2024.

[31] T. M. Team, "Modelscope: bring the notion of model-as-a-service to life." https://github.com/modelscope/modelscope, 2023.

[32] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," 2024.

[33] T. Fehrer, A. Egger, D. Chvirova, J. Wittmann, N. Wördehoff, W. Kratsch, and M. Röglinger, "Business Processes in IT Asset Management Multimedia Event Log," 2024.

[34] A. Gavric, D. Bork, and H. Proper, "Enriching business process event logs with multimodal evidence," in *The 17th IFIP WG 8.1 Working Conference on the Practice of Enterpris Modeling (PoEM)*, 2024.

[35] K. Lee, D. Ognibene, H. J. Chang, T.-K. Kim, and Y. Demiris, "Stare: Spatio-temporal attention relocation for multiple structured activities detection," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5916–5927, 2015.