Petri Net Structure-Driven Video Generation

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

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

Henderik A. Proper
TU Wien
Vienna, Austria
henderik.proper@tuwien.ac.at

Abstract

Recent advances in video generation have unlocked new opportunities for simulating real-world activities. Yet, existing models often struggle to faithfully represent structured, multi-step processes—such as those found in business workflows—resulting in temporally inconsistent or semantically incoherent outputs.

To address this gap, we propose *Petri Net Structure-Driven Video Generation*, an approach that leverages formal process models to guide the generation of coherent and semantically grounded video simulations. Specifically, we incorporate process-aware structural information through: (i) domain-specific prompting enriched with process semantics, (ii) storyboard construction using reference frames extracted from real-world process evidence, and (iii) synthetic reference frames informed by Petri Net structures.

We evaluate our method across multiple domains and show that grounding generation in process model structure improves temporal coherence, semantic fidelity, and user-perceived realism. Our approach demonstrates how structured symbolic representations can enhance generative video systems, opening new directions for process-aware visual synthesis.

1 Introduction

Business process simulation (BPS) has long served as a tool for understanding, analyzing, and optimizing organizational workflows [Mendling et al., 2013]. Developing simulation models manually is a labor-intensive and error-prone process, filled with numerous challenges [Van Der Aalst, 2014]. 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 [Camargo et al., 2022], quality evaluation frameworks [Chapela-Campa et al., 2023], runtime integration strategies [Meneghello et al., 2023], and agent-based discovery frameworks [Kirchdorfer et al., 2024]. Traditional simulation techniques primarily focus on replaying event logs [Van Der Aalst, 2014]. However, the emergence of video generation models 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* [Saunders et al., 2018, Gagliano, 1988].

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% [Fortune Business Insights, 2025]. Furthermore, recent research has shown how conceptual models can be transformed

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

into multimodal outputs (images and audio) thereby paving the way for video-based business process simulations by facilitating that state-of-the-art video generation models offer **the fusion of textual and visual cues** during the video generation [Liu et al., 2024].

Despite these advances, current video generation approaches often fall short in capturing the *inherent dynamics* and *structural complexity* of **business processes** [Gavric et al., 2024d]. 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.

The remainder of this paper is organized as follows. Section 2 reviews background for understanding discovery of business process simulations and multimodal evidence integration within business process management, in regard to the video generation. Section 3 describes our approach for the development of video generation instructions from a straight-forward prompting to the proposed Petri Net structure-driven video generation methodology. Section 4 presents the evaluation and its results. Finally, Section 5 concludes the paper.

2 Background

This work bridges *structure-driven* text-to-video diffusion, *business process simulation* (BPS) discovery, and *multimodal* business process management (BPM); we provided an extended survey in Appendix A. Structure-aware video generators use pose/depth/layout and graph cues to improve temporal and interaction fidelity [Wang et al., 2023a,b, Li et al., 2021, Zhou et al., 2023, Zheng et al., 2023], with evaluation moving beyond aesthetics to structure-consistency metrics [Huang et al., 2024, Han et al., 2025]. BPS discovery spans control-flow-first and resource-first (agent-based) paradigms, with learning-based accuracy benchmarks and work on trust and context [Rozinat et al., 2009, Kirchdorfer et al., 2024, Camargo et al., 2022, Chapela-Campa et al., 2023]. In BPM, large video models and LLM prompting enable process visualization and analysis from multimodal evidence [Lin et al., 2024, Liu et al., 2024, Kratsch et al., 2022, Gavric et al., 2024b], which we leverage via domain-rich prompts, operational images, and state-aware interpolation for structurally faithful video simulations.

To formalize process dynamics, we adopt *Petri nets* as our underlying representation. Petri nets [Petri and Reisig, 2008] are a well-established mathematical modeling language for concurrent, distributed, and resource-sensitive systems. They provide a bipartite graph structure of *places* (representing conditions or resources) and *transitions* (representing events or activities), connected through directed arcs. Their token-based execution semantics naturally capture causality, concurrency, and choice—key aspects of business processes. In the following, we provide a formal definition that grounds our approach.

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.

3 Video Business Process Simulations

In this section, we outline our methodology. We first refer to the video generation and how one can integrate process models into such process in Appendix B. To complement the methodological overview, we provide in Appendix C a detailed explanation of the our video generation strategies. In the following, we state their *formal definitions*, which specify the input components, transformations, and resulting storyboards that constitute each approach.

Approach A. 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. Video generation tool 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.

Approach B. 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, video generation tool realizes frame interpolation mechanism $\mathcal{F}_{interp}$: $S_B = \mathcal{F}_{interp}(S_{key})$. By using real operational images, Approach B grounds the simulation in authentic process visuals thereby enhancing the contextual relevance of the generated video.

Approach C. 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.

Hybrid Approach. 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.

We detail prototyping of those approaches for video business process simulation generation denoted as *Petri Net Structure-Driven Video Generation Guidance* in Appendix D.

4 Evaluation

We present results of our multi-faceted evaluation for the proposed video-generation approaches, conducted as described in Sec. E.

The 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.

Results. The evaluation agents 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 evaluation agents 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 .

Answer to RQs. 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 .

5 Conclusion

In this paper, we analyzed possibilities of bridging the gap between traditional process mining and modern video-generation capabilities, 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

Manuel Camargo, Marlon Dumas, and Oscar 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*, pages 286–302. Springer, 2019.

Manuel Camargo, Marlon Dumas, and Oscar González-Rojas. Automated discovery of business process simulation models from event logs. *Decision Support Systems*, 134:113284, 2020.

Manuel Camargo, Marlon Dumas, and Oscar 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*, pages 55–71. Springer, 2022.

David Chapela-Campa, Ismail Benchekroun, Opher Baron, Marlon Dumas, Dmitry Krass, and Arik Senderovich. Can i trust my simulation model? measuring the quality of business process simulation models. In *International Conference on Business Process Management*, pages 20–37. Springer, 2023.

Tobias Fehrer, Andreas Egger, Diana Chvirova, Jakob Wittmann, Niklas Wördehoff, Wolfgang Kratsch, and Maximilian Röglinger. Business Processes in IT Asset Management Multimedia Event Log, 2024.

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. URL https://www.fortunebusinessinsights.com/ai-video-generator-market-110060.

Martha E Gagliano. A literature review on the efficacy of video in patient education. *Academic Medicine*, 63(10):785–92, 1988.

Aleksandar Gavric, Dominik Bork, and Henderik 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)*, 2024a.

Aleksandar Gavric, Dominik Bork, and Henderik Proper. Multimodal process mining. In *26th International Conference on Business Informatics (CBI)*, 2024b.

Aleksandar Gavric, Dominik Bork, and Henderik Proper. 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)*, 2024c.

Aleksandar Gavric, Dominik Bork, and Henderik Proper. How does uml look and sound? using ai to interpret uml diagrams through multimodal evidence. In *43rd International Conference on Conceptual Modeling (ER)*, 2024d.

Thomas Grisold, Han van der Aa, Sandro Franzoi, Sophie Hartl, Jan Mendling, and Jan Vom Brocke. A context framework for sense-making of process mining results. In *2024 6th International Conference on Process Mining (ICPM)*, pages 57–64. IEEE, 2024.

Michal Halaška and Roman Šperka. Is there a need for agent-based modelling and simulation in business process management. *Organizacija*, 51(4):255–269, 2018.

Hao Han et al. Video-bench: Human-aligned video generation benchmark. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Han_Video-Bench_Human-Aligned_Video_Generation_Benchmark_CVPR_2025_paper.pdf.

Xinting Hu, Haoran Wang, Jan Eric Lenssen, and Bernt Schiele. Personahoi: Effortlessly improving personalized face with human-object interaction generation. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Hu_PersonaHOI_Effortlessly_Improving_Face_Personalization_in_Human-Object_Interaction_Generation_CVPR_2025_paper.pdf.

Junchao Huang, Xinting Hu, Zhuotao Tian, Shaoshuai Shi, and Li Jiang. Edit360: 2d image edits to 3d assets from any angle. *arXiv preprint arXiv:2506.10507*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Yuming Jiang, Kelvin C.K. Chan, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Chenyang Si, Ziwei Liu, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, 2024.

Nicholas R. Jennings, Peyman Faratin, MJ Johnson, Timothy J. Norman, P O'brien, and Mark E Wiegand. Agent-based business process management. *International Journal of Cooperative Information Systems*, 5(02n03):105–130, 1996.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2023.

Lukas Kirchdorfer, Robert Blümel, Timotheus Kampik, Han Van der Aa, and Heiner Stuckenschmidt. Agentsimulator: An agent-based approach for data-driven business process simulation. In *2024 6th International Conference on Process Mining (ICPM)*, pages 97–104. IEEE, 2024.

Wolfgang Kratsch, Fabian König, and Maximilian Röglinger. Shedding light on blind spots–developing a reference architecture to leverage video data for process mining. *Decision Support Systems*, 158:113794, 2022.

Anna Kukleva et al. Taec: Unsupervised action segmentation with temporal-aware embedding and clustering. In *CEUR Workshop Proc.*, 2023. URL https://ceur-ws.org/Vol-3527/short5.pdf.

Kyuhwa Lee, Dimitri Ognibene, Hyung Jin Chang, Tae-Kyun Kim, and Yiannis Demiris. Stare: Spatio-temporal attention relocation for multiple structured activities detection. *IEEE Transactions on Image Processing*, 24(12):5916–5927, 2015.

Siyao Li, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, 2022.

Yuezun Li, Jing Lin, Zekun Wang, Heng Ding, Lingxi Xie, Qi Tian, and Jiebo Luo. Ag2vid: End-to-end compositional video synthesis from action graphs. In *ACM MM*, 2021.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Guohai Lin et al. Replace anyone in videos. *arXiv preprint arXiv:2409.19911*, 2025. URL https://arxiv.org/abs/2409.19911. v2 May 2025.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Jan Mendling, Hajo A Reijers, Marcello La Rosa, and Marlon Dumas. Fundamentals of business process management. In *GI-Jahrestagung*, page 157. Springer, 2013.

Francesca Meneghello, Chiara Di Francescomarino, and Chiara Ghidini. Runtime integration of machine learning and simulation for business processes. In *2023 5th International Conference on Process Mining (ICPM)*, pages 9–16. IEEE, 2023.

Ali A Mohamed and Brandon Lucke-Wold. Text-to-video generative artificial intelligence: sora in neurosurgery. *Neurosurgical Review*, 47(1):272, 2024.

Julian Neuberger, Lars Ackermann, Han van der Aa, and Stefan Jablonski. A universal prompting strategy for extracting process model information from natural language text using large language models. In *International Conference on Conceptual Modeling*, pages 38–55. Springer, 2024.

Carl Adam Petri and Wolfgang Reisig. Petri net. *Scholarpedia*, 3(4):6477, 2008.

Adrian Rebmann, Fabian David Schmidt, Goran Glavaš, and Han van Der Aa. Evaluating the ability of llms to solve semantics-aware process mining tasks. In *2024 6th International Conference on Process Mining (ICPM)*, pages 9–16. IEEE, 2024.

Anne Rozinat, Ronny S Mans, Minseok Song, and Wil MP van der Aalst. Discovering simulation models. *Information systems*, 34(3):305–327, 2009.

Alicia F Saunders, Fred Spooner, and Luann Ley Davis. Using video prompting to teach mathematical problem solving of real-world video-simulation problems. *Remedial and Special Education*, 39(1): 53–64, 2018.

Step-Video Team. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. URL https://arxiv.org/abs/2502.10248.

Emilio Sulis and Kuldar Taveter. *Agent-Based Business Process Simulation*. Springer, 2022.

Andrei Tour, Artem Polyvyanyy, and Anna Kalenkova. Agent system mining: vision, benefits, and challenges. *IEEE Access*, 9:99480–99494, 2021.

Andrei Tour, Artem Polyvyanyy, Anna Kalenkova, and Arik Senderovich. Agent miner: An algorithm for discovering agent systems from event data. In *International Conference on Business Process Management*, pages 284–302. Springer, 2023.

Wil MP Van Der Aalst. Business process simulation survival guide. In *Handbook on business process management 1: Introduction, methods, and information systems*, pages 337–370. Springer, 2014.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *NeurIPS*, 2023a. URL https://videocomposer.github.io/.

Zhicai Wang et al. Enhance image classification via inter-class image mixup with diffusion model. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Wang_Enhance_Image_Classification_via_Inter-Class_Image_Mixup_with_Diffusion_Model_CVPR_2024_paper.pdf.

Zhouxia Wang, Yixin Yao, Yixiao Li, Yifan Liu, Xintao Wang, Ying Shan, Yu Qiao, et al. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023b. URL https://arxiv.org/abs/2312.03641.

Yongliang Wu, Xinting Hu, Yang Xu, et al. Unlearning concepts in diffusion model via concept domain correction. *AAAI*, 2025. doi: 10.1609/aaai.v39i8.32917. URL https://dl.acm.org/doi/10.1609/aaai.v39i8.32917.

Yikang Zheng, Hongjie Yan, Siyao Li, Sheng Zhang, Guangrun Chen, Ziwei Liu, and Chen Change Loy. Panoptic video scene graph generation. In *ICCV*, 2023.

Yufan Zhou, Liang Zhang, Yujun Shi, Yandong Guo, Baoyuan Wang, et al. Scene graph guided video generation with diffusion models. In *ICCV*, 2023.

A Related Work

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

A.1 Video Generation

Structure-driven controls for video diffusion. Modern text-to-video diffusion increasingly supports *explicit structure* as control signals, such as human pose, depth, segmentation, trajectories, and camera motion. Compositional control in VideoComposer unifies textual, spatial, and temporal conditions and exploits motion vectors for temporal guidance, improving inter-frame consistency and motion controllability [Wang et al., 2023a]. MotionCtrl further disentangles and jointly controls object and camera motion via a unified controller applicable across popular video backbones [Wang et al., 2023b]. Recent training-free pipelines focus on *localized human* control in-the-wild scenes (e.g., replacement/insertion while preserving motion cues) [Lin et al., 2025]. Beyond motion fields, layout- and graph-based controls leverage high-level scene structure. Action Graph-to-Video casts activities as compositional graphs that drive generation [Li et al., 2021], and scene-graph guided video synthesis conditions diffusion on structured object relations for coherent spatio-temporal interactions [Zhou et al., 2023]. Relatedly, panoptic video scene-graph generation (PVSG) advances structured *understanding* of entities, relations, and dynamics [Zheng et al., 2023], feeding richer constraints into generative pipelines.

Pose- and motion-centric priors. A complementary line learns strong priors for human motion. MotionGPT treats motion as a "language," enabling text-driven motion generation, captioning, and prediction in a unified framework [Jiang et al., 2023]. Music-conditioned motion generation likewise exploits discrete motion units and long-horizon structure [Li et al., 2022]. These priors naturally align with pose/keypoint controls used in structure-driven video synthesis.

Editing, identity consistency, and 360° control. Structure-driven objectives are also realized through editing frameworks that preserve identity and global geometry while changing content. PersonaHOI demonstrates training/tuning-free identity-consistent human–object interaction generation by coupling a personalized-face diffusion branch with HOI-guided generation [Hu et al., 2025]. Edit360 lifts 2D edits to multi-view-consistent 3D asset edits *via* video diffusion backbones and

anchor-view propagation, enabling any-angle consistency [Huang et al., 2025]. Large open models such as Step-Video-T2V systematize foundation-model practice (Video-VAE compression, DiT with flow matching, video-DPO), offering stronger backbones for controllable/structured pipelines [Step-Video Team, 2025].

Learning signals, unlearning, and safety. As structure becomes richer, controlling *what a model should not generate* is likewise important. Concept-domain correction enables concept unlearning in diffusion while preserving model utility [Wu et al., 2025]. Orthogonally, diffusion-based mixup augments discriminative training with generative priors, hinting at tighter loops between control signals and robust recognition [Wang et al., 2024].

Temporal segmentation and activity structure. Unsupervised/weakly supervised action segmentation provides *temporal scaffolds* (ordering constraints, state boundaries) that structure-driven generators can exploit. For instance, TAEC introduces temporal-aware embeddings and clustering for unsupervised action segmentation [Kukleva et al., 2023], aligning well with storyboarded or graph-driven synthesis.

Evaluation and benchmarks. Evaluating structure adherence requires beyond-aesthetics metrics. VBench decomposes video quality into disentangled dimensions (e.g., identity consistency, motion smoothness, spatial relations) [Huang et al., 2024], while newer suites emphasize intrinsic faithfulness and human alignment [Han et al., 2025]. Such benchmarks are increasingly adopted to quantify structural fidelity in controllable video generation.

A.2 Discovering Business Process Simulation

In discovering business process simulation models [Rozinat et al., 2009], 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 [Kirchdorfer et al., 2024], 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. [Jennings et al., 1996] laid the foundation for agent-based BPM nearly 30 years ago. Later, [Halaška and Šperka, 2018] assessed its need, [Sulis and Taveter, 2022] advanced simulations, [Tour et al., 2021, 2023] mined agent systems, and [Kirchdorfer et al., 2024] introduced a resource-first version.

Research on learning accurate representation of BPS models [Camargo et al., 2019] and on the automated discovery of BPS models from event logs [Camargo et al., 2020, 2022], has set early benchmarks in BPS model discovery accuracy. Additionally, the authors in [Chapela-Campa et al., 2023, Grisold et al., 2024] have further addressed the challenges of assessing and contextualizing BPS models. 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.

A.3 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 [Lin et al., 2024]. Both, open-source [Lin et al., 2024], and closed-source trained models [Liu et al., 2024], are introduced and benchmarked, showcasing promising applications even in the domain of neurosurgery [Mohamed and Lucke-Wold, 2024]. In parallel, universal prompting strategies have been explored and evaluated the capabilities of large language models for semantics-aware process mining tasks [Neuberger et al., 2024, Rebmann et al., 2024]. Explorations into the integration of machine learning with simulation [Camargo et al., 2019, 2022, Meneghello et al., 2023] 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, [Kratsch et al., 2022, Gavric et al., 2024b] for process discovery from multimodal data, and [Gavric et al., 2024d,c], 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.

B Video Generation

For video generation, we employ a video generation engine, OpenAI's SORA[1]. 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 image; or by (*B*) *synthesizing a storyboard* which serves as the blueprint for generating a continuous video output.

Despite their capabilities, video generation tools currently operates on a primarily generic framework. Its design, while powerful for a broad range of applications [Liu et al., 2024, Mohamed and Lucke-Wold, 2024], 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.

To overcome these limitations, we propose enhancing video generation tools 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 [Saunders et al., 2018, Gagliano, 1988].

C 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.

C.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. 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).

[1] `https://openai.com/sora/`

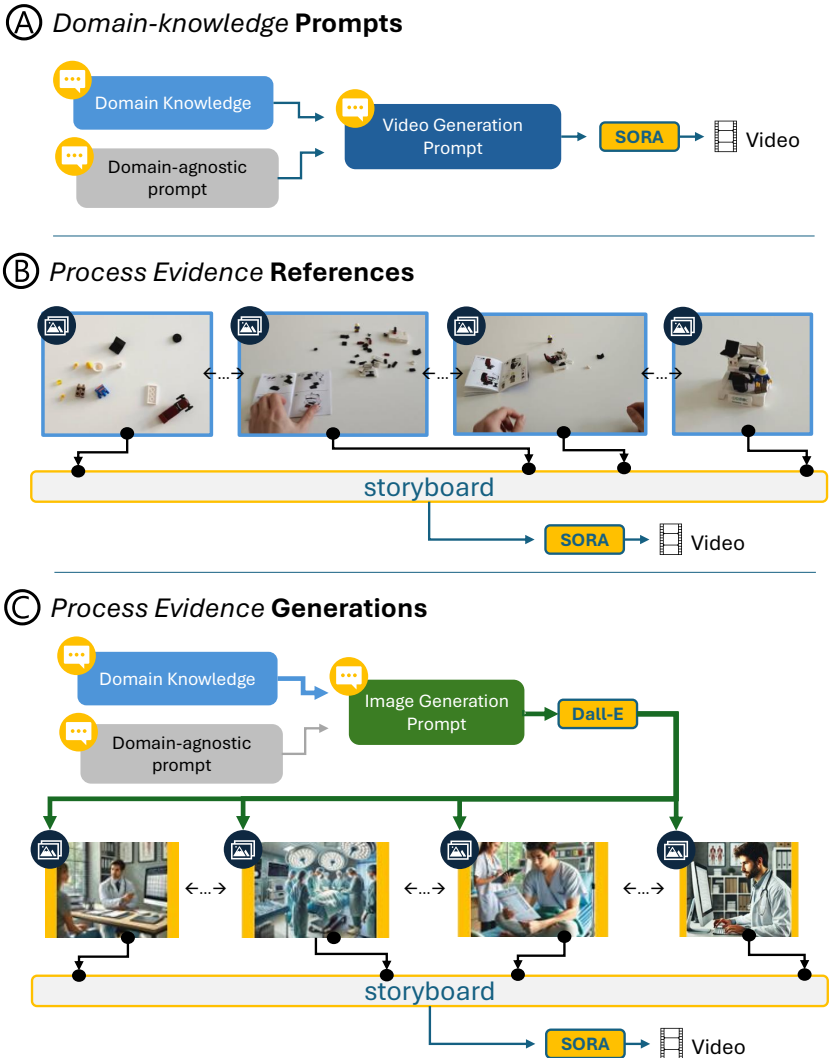

Figure 1: **Overview of the proposed Approaches.** (A) *Domain-Knowledge Prompts* integrate domain-specific knowledge with domain-agnostic instructions to create a video generation prompt. (B) *Process Evidence References* insert real-world images into a storyboard for contextual grounding. (C) *Process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

C.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 the actual business process execution. These references serveto video generation tool as keyframes to anchor the storyboard, ensuring visual fidelity to the underlying process.

C.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.

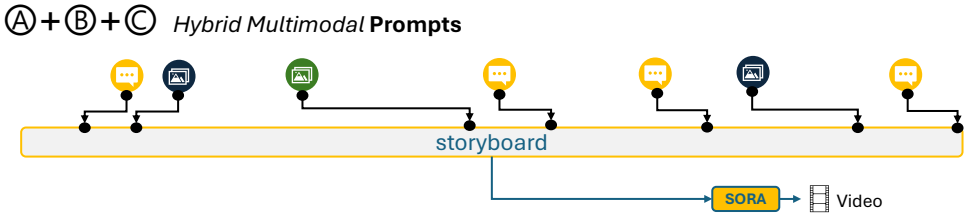

Figure 2: **Hybrid Multimodal Prompts (A + B + C).** In the hybrid approach, we merge domain-knowledge prompts (A), process evidence references (B), and process evidence generations (C) into a single, multimodal storyboard that is then processed by a video generation tool to yield the final video simulation.

C.3.1 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 multimodal 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).

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.

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.

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.

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.

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

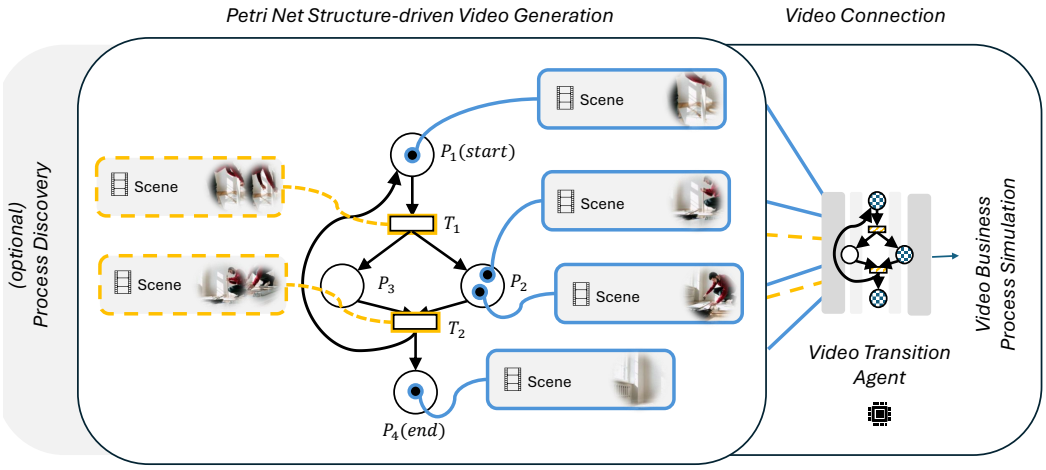

Figure 3: **Petri Net Structure-Driven Video Generation Architecture.** An optionally discovered Petri Net with places $(P_1, \ldots, P_4)$ and transitions $(T_1, T_2, \ldots)$. Each place corresponds to one or more *Scene(s)* in the final video. The *Video Transition Agent* then navigates through the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

$\{T_1, T_2, \ldots\}$, 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.

E Evaluation Setup and Discussion

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 [Fehrer et al., 2024], (2) DNA Testing [Gavric et al., 2024a], (3) Cooking [Kratsch et al., 2022] (which uses data from [Lee et al., 2015]), and (4) IKEA [Gavric et al., 2024b]. 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 our video generation tool training data.

Options Generation. For each domain, A single reference Petri Net was obtained per domain from several datasets [Fehrer et al., 2024, Gavric et al., 2024a, Kratsch et al., 2022, Gavric et al., 2024b] for E1-E4, created using multimodal process discovery tool [Gavric et al., 2024a] 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 [Kirchdorfer et al., 2024] 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.

E.1 User Study

A total of 50 ChatGPT 4o agents were prompted for the study. For each video in five (E1–E5) evaluation domains, the agents 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, agents 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), agents 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 agents 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*).

E.2 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 a video generation tool.

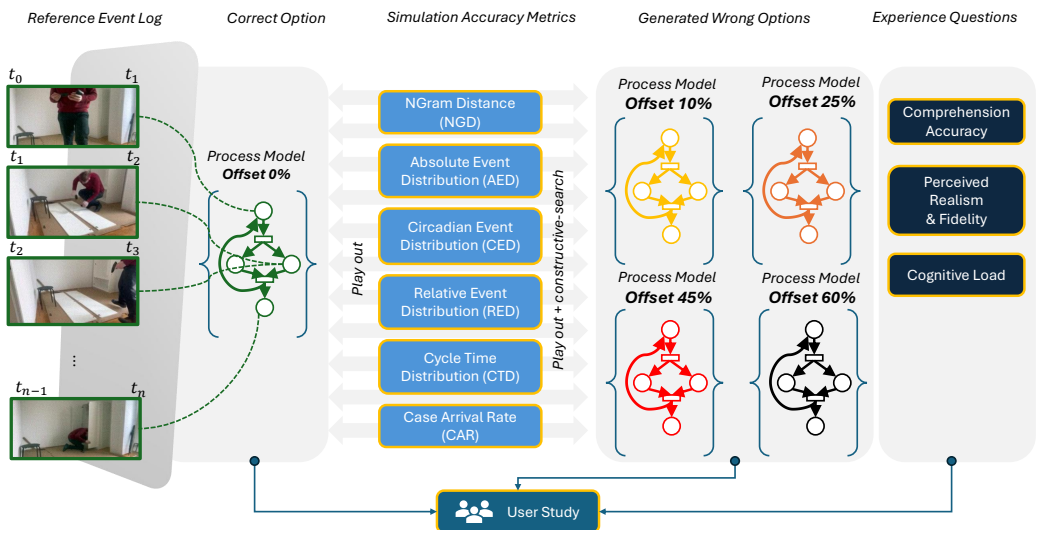

Figure 4: Illustration of the evaluation pipeline.

E.3 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.

E.4 Threats to Validity and Limitations.

Internal validity may be affected by participant biases in the user study, such as prior familiarity with process modeling concepts. External validity is constrained by the specific evaluation domains chosen—while our datasets span multiple business processes (five domains), the generalizability of our findings to other domains, especially those requiring ultra-fine visual detail (e.g., complex mechanical assemblies), real-life production-grade video training quality, consistency in longer videos (currently just 20s) with more actors and objects in use, remains uncertain. All these areas deserve research on their own.

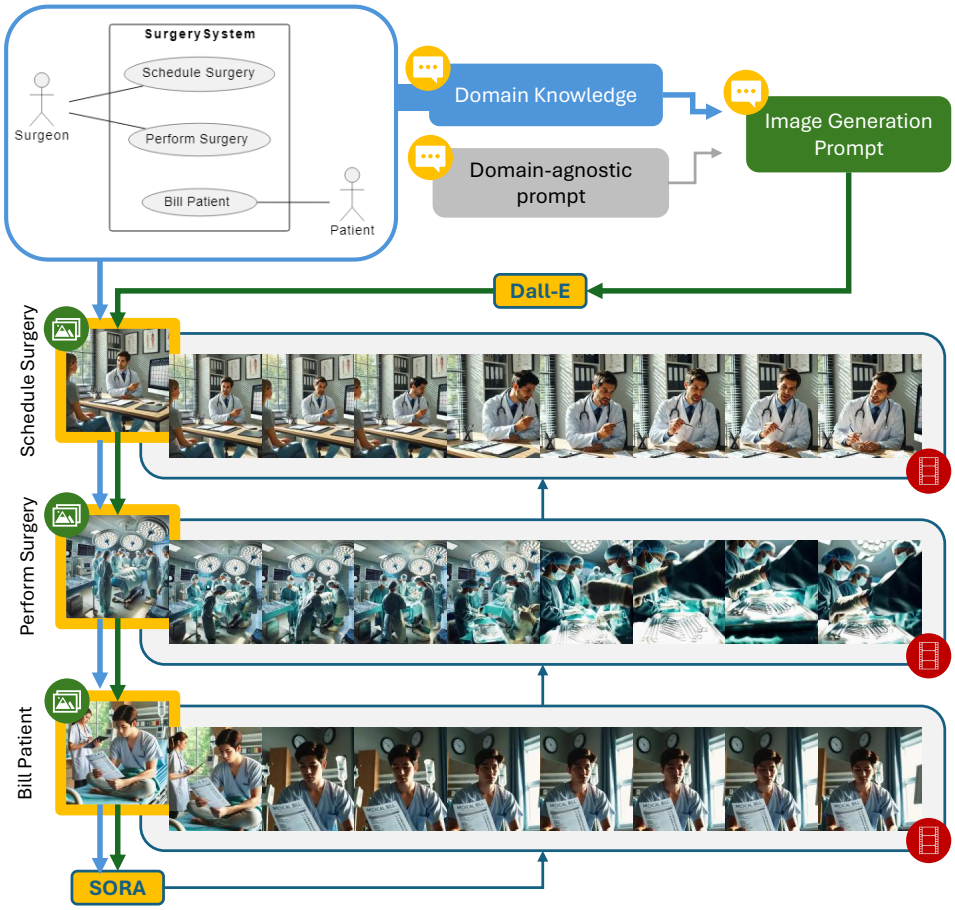

Figure 5: **Example application of Approach C in a surgical domain.** Domain knowledge ("Surgery System") and domain-agnostic prompts (general instructions) guide an image generation model (DALL-E). Synthetic images for each major step ("Schedule Surgery," "Perform Surgery," "Bill Patient") are then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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly articulate the core contributions and scope of the paper. The main claims—regarding the proposed method, its theoretical motivation, and the empirical improvements shown—are consistent with what is actually delivered and evaluated in the rest of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A discussion of limitations is provided.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: While intuition and proof sketches are given in the main text, full formal proofs and all assumptions are deferred to future work. The paper does not yet include a fully rigorous appendix with all derivations.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The paper provides details about datasets, metrics, and evaluation procedures, but not all hyperparameters and preprocessing steps are documented. Thus, exact reproduction would be difficult without additional supplementary information.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: code and data are not released to preserve anonymity. We plan to release both upon acceptance with sufficient instructions, but at this stage reviewers cannot directly access them.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: The paper specifies major architectural choices, optimizers, and datasets, but omits full hyperparameter sweeps, random seed management, and some environment details. These will be included in supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Results are reported as mean performance across runs but without formal confidence intervals or statistical tests. Variability is mentioned qualitatively, but not quantified with error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: We note that experiments were run on GPUs but do not provide detailed estimates. More precise resource accounting would be required for reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: No ethical violations are present.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We reflect on both potential positive applications and negative risks We outline possible mitigation strategies.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release high-risk models or datasets, so no special safeguards are required.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external datasets and code bases used in the experiments are properly cited, and their licenses (e.g., CC-BY, MIT) are respected and stated where applicable.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: This paper does not introduce new datasets or software packages as standalone assets, so structured documentation is not applicable.

    Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The paper does not involve human-subject studies or crowdsourcing tasks, hence no such details are included.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The work does not involve human subjects and therefore no IRB approval was required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were not used beyond writing support—e.g., in generating synthetic baselines, evaluation heuristics, or methodological components—and clarify their role in the research pipeline.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.