Comics as Process Model Notation: Blending Object-Centric Event Logs and Multimodal Data in Visually Enhanced Narratives

Process modeling is essential for understanding, analyzing, and improving organizational workflows, yet conventional notations such as BPMN often present steep learning curves and limited engagement for non-expert stakeholders. To address these shortcomings, we introduce a novel paradigm that treats comics as a first-class notation for process models. We present a framework, *Visually Enhanced Narratives* (ViEnNa), that defines a domain-specific language in which each comic panel represents a process step augmented with object-centric event-log semantics and linked multimodal evidence (images, video, audio, text). Our evaluation demonstrates that ViEnNa enhances the perceived appeal of process models—even surpassing non-traditional approaches such as SAP Scenes and AI-generated non-ViEnNa comics—while preserving the modeling accuracy typical of established notations like BPMN and its object-centric variant.

Keywords: *AI-assisted process modeling; Semantics in process modeling; Generative AI; Visual modeling language; Multimodal data; Object-centric event logs.*

1 Introduction

Understanding and communicating business processes is a foundational activity in information systems engineering, supporting tasks such as process improvement, compliance checking, and system design (van der Aalst and Carmona 2022). Traditional process models, often depicted as flowcharts or diagrams, aim to capture the sequence and logic of activities within a process. However, these models frequently fall short of representing the complexities of real-world processes. They often lack contextual information, omit the diversity of data sources involved, and fail to capture the dynamic, interactive nature of many processes (Bardes et al. 2024). To address these shortcomings, researchers have explored various avenues to enhance process modeling. Object-centric event logs (OCEL), and more recently, OCEL 2.0 (Berti et al. 2024a), have emerged as a standard to capture the granular details of process execution, depicting changes in objects, focusing on the interactions between objects within a process, and qualifying relationships to other objects or specific events.

Concurrently, the growth of digital surveillance (Grisold et al. 2024) and multimodal data sources, including images, videos, audio recordings, and sensor data, offers a wealth of information that can be used for the discovery of process models (Gavric et al. 2024b). Integrating such data into process rep-

resentations can provide a more inclusive (Lukyanenko et al. 2023) and more realistic view of process execution. Furthermore, the rise of artificial intelligence (AI), and its influence on conceptual modeling, as shown by Bork et al. (2023); Fill et al. (2023); Feuerriegel et al. (2024), opens new possibilities for analyzing and interpreting these diverse data sources.

We aim to blend object-centric event logs with multimodal real-world data while preserving the intuition of human-model interaction (**?**). To make process models more accessible to a broader audience, including those who are not domain modeling experts, we propose exploring the use of *comic-like process models*. Comics are increasingly explored through computational methods (Laubrock and Dunst 2020; Nguyen et al. 2019), while some of them use an ontology to assist automated interpretation (Guérin et al. 2017). Comics have proven effective in science communication (Bucher and Boy 2018), legal contexts through *contracts that have and give voice* (Ketola et al. 2024), and even source-code visualization (Heidrich et al. 2024). Recent work shows that AI can generate comics automatically (Kumar et al. 2024), further positioning them as a viable medium for multimodal process representation. Our ambition is to improve human-model interaction to not only focus on how models are represented but also embedding the model in real-world background data/evidence.

As such, the research objective of this paper is defined as follows: **(RO)** *Design a modeling framework for comic-like representations of processes that enhances the perceived appeal of process models compared to existing approaches, while maintaining comprehension accuracy.* Intended users include process modeling experts, non-experts, and AI agents, for both creating and interpreting process models. The challenge is not merely to visualize multimodal data pulled from memory units but to understand and synthesize it in a way that leads to informed, evidence-aware decisions in process modeling. This capability enables the use of visual metaphors and object- or character-driven narratives in a way that is aware of the semantics of multimodal data inputs. We present a process modeling framework with the umbrella term ***Vi**sually **En**hanced **Na**rratives* (**ViEnNa**) that encompasses a modeling language, a method, and a tool. ViEnNa models are process models, enriched with and anchored to visual elements to contextualize narrative process flows, integrate the storyline logic embedded within the models and evidence data, and present it in a format that aims to be intuitively comprehensible to non-process modeling experts.

The remainder of this paper is structured as follows. Section 2 reviews the existing literature and situates our work within the broader context of related research, highlighting how our contributions build upon and differ from previous studies. Section 3 introduces and describes the ViEnNa framework. We provide the derivation of the ViEnNa process modeling language in Section 4. Section 5 presents a thorough evaluation of our framework, including the methods used for assessment and the results obtained. Section 6 discusses implications from the evaluation. Finally, Section 7 concludes the paper.

The created tool, datasets, models, and evaluation details are available as supplementary material at: `https://anonymous.4open.science/r/viennacomics/`.

2 Related Work

In this research, we perceive process models as a *design of events* that structures the story and serves as the backbone of visually enhanced narratives. This perspective aligns with the view that storytelling is not merely the sequencing of information, but a deliberate orchestration of events toward meaning:

> "A story is not an accumulation of information strung into a narrative, but a design of events to carry us to a meaningful climax."
>
> — *Robert McKee, Storylogue*

Consequently, our related work section is structured around two key themes: (i) Transition from conventional to non-conventional process modeling, and (ii) Automated process modeling and comics creation. To provide a structured overview, we grouped existing contributions according to two core criteria: (C1) **Representation Formality**—ranging from symbolic and formal to visual and narrative-based representations, and (C2) **Automation and Human-Centricity in Modeling**—spanning automated, AI-driven techniques to participatory, user-centered approaches. For each group, we highlight relevant approaches.

2.1 Transition from Conventional to Non-Conventional Process Modeling

Since the early software crisis (Naur et al. 1969), researchers have sought to solve real-world problems such as hiring, production planning, and service delivery through computing and software-based approaches. Problem-oriented yet formal methods emerged, leading to the concept of *modeling*, which involves the creation of comprehensive models that integrate data, behavior, structure, and actors—both human and technical—into cohesive representations Fettke and Reisig (2020). Frameworks such as Business Process Model and Notation (BPMN) 2.0 (Object Management Group 2013), and Directly-Follows Graphs (DFG) (van der Aalst and Carmona 2022) have since become standards for process modeling, using formal, graphical notations. Object-Centric Event Logs (Berti et al. 2024a) extend this paradigm by capturing interactions among multiple object types within a process and enabling object-centric (oc-) process models that provide more granular process perspectives. [(C1) Symbolic, Formal]

However, conventional modeling approaches often fall short in terms of usability and accessibility, particularly for non-expert stakeholders. This has led to growing interest in non-conventional modeling methods that prioritize interpretability and engagement. Notable examples include *spatial conceptual modeling* (Fill 2024), which uses augmented reality to anchor abstract process logic in the physical environment [(C2) Spatial, Visual], and *Domain Storytelling* (Hofer et al. 2021), which facilitates participatory modeling by bringing together developers and domain experts [(C2) Participatory, Narrative]. As a relevant example of non-conventional process models used in industry, we highlight SAP Scenes (SAP 2019) that represent a storytelling-based modeling approach using character-driven scenarios to support shared understanding and haptic feeling [(C2) Haptic, Professional Training], similar to Miron et al. (2019). To formalize such approaches, Domain-Specific Modeling Language (DSML) is often employed. Frameworks like those by Frank (2013); Jannaber et al. (2017); Karsai et al. (2014), that guide the structured design of DSMLs and the integration of domain knowledge into visual syntax and metamodeling concepts [(C1) Domain-Driven, Formalization]. We also consider comics as a non-conventional model-

ing approach. Comics provide a visually enhanced narrative structure that is already well-understood in domains such as law (Ketola et al. 2024), software engineering (Heidrich et al. 2024), and science communication (Bucher and Boy 2018), yet their use as a process model within the *business process management* (BPM) domain stays unexplored [C2 Engaging, Informal].

2.2 Automated Process Modeling and Comics Creation

Automated process modeling, particularly in the context of object-centric process mining, has gained traction in recent years. Van der Aalst and Berti (2020) proposed a method for discovering object-centric Petri nets from OCELs, enabling expressive modeling of intertwined object behaviors [C1 Formal, Data-aware, Automated]. Rebmann et al. (2022) presented a method to derive object-centric logs from traditional event logs through semantic and control-flow analysis. Other advances include data-aware extensions (Goossens et al. 2022), predictive modeling (Rohrer et al. 2022), conformance checking (Xiu and Li 2023), and quality assurance in discovered models (Benzin et al. 2023) [C1 Towards Full Automation]. Complementing these developments are human-centered and AI-supported approaches to visualization and model creation. One notable work, Scene2Model (Völz et al. 2024) makes automated object description and attribute generation, along with comprehensive scene summaries [C1 Automated, C2 AI-Supported, Visual Analysis] working with non-conventional process models. Work such as that of Fill and Muff (2023) explores how large language models can be prompted to produce structural visualizations [C2 LLM Prompting, Structural Abstraction]. Similarly, Gavric et al. (2024a) propose augmenting business process event logs with multimodal evidence for improved fidelity, bridging visual, textual, and structural data streams into unified representations [C1 Evidence-Driven, C2 Multimodal].

2.3 Motivation and Our Position

Missing gap. Despite their formal rigor and widespread adoption, conventional modeling techniques such as BPMN (Object Management Group 2013) and DFG (van der Aalst and Carmona 2022) often suffer from key limitations when applied in real-world settings (Haisjackl et al. 2018): their reliance on abstract syntax and domain-specific terminology can create steep learning curves for non-technical stakeholders. Non-conventional approaches, like spatial conceptual modeling (Fill 2024), offer greater inclusivity and interactivity, but they still face challenges in terms of more intuitive, multimodal, and image synthesizing-driven representations of processes. Automated process modeling streams of research underscore the potential—but also current limitations—of AI-assisted process modeling, as shown in Muff and Fill (2024). While C1-marked contributions ensure syntactic precision, they often sacrifice narrative engagement and intuitive accessibility. Conversely, C2-marked contributions boost engagement but lack the perceived appeal offered by image-synthesizing integration within the narrative structure. Existing modeling paradigms inadequately balance these dimensions.

Our position. We position ViEnNa as a new form of non-conventional process modeling, built as a DSML to represent processes in the form of comics. Narratives have long been a subject of systematic analysis (Esin 2011), and theories like Bateman's multimodal discourse framework (Bateman and Wildfeuer 2014b) provide foundations for interpreting comic-based communication. Advances in com-

putational analysis of comics (Laubrock and Dunst 2020; Nguyen et al. 2019; Guérin et al. 2017; Tanaka et al. 2007; Kim et al. 2024) and the growing use of AI in automatic comic generation (Kumar et al. 2024; Charles et al. 2024; Chen and Jhala 2024) further support this trajectory. While comics improve accessibility and narrative engagement, open questions remain regarding their integration with formal conceptual modeling logic. ViEnNa addresses this challenge by providing a structured yet expressive modeling language for object-centric, multimodal processes—aiming to boost engagement through the integration of image synthesis within the narrative flow.

3 ViEnNa Modeling Framework

The ViEnNa modeling framework is designed with built-in AI capabilities that enable multimodal data inputs to represent (real-world) data within the process model.

Our Methodology. Our methodology blends structured event data with multimodal evidence to synthesize process models in the form of comic panels. This is achieved by integrating three core components: pixel-level image generation using diffusion-based models (cf. Sect. 3.1), multimodal data processing via AI-supported retrieval mechanisms (cf. Sect. 3.2), and, most importantly, a novel domain-specific modeling language tailored for narrative-first process representation, whose derivation is provided in Section 4.

Preliminaries. Our methodological foundation builds upon the formalism of object-centric event data representation Berti et al. (2024a), which serves as a structured backbone for mapping raw evidences to high-level process elements.

Definition 1 (Process Comic) *A* Process Comic *is a visual process representation defined as a tuple* $\mathscr{C} = \langle F, O, A, L \rangle$*, where* F *is a sequence of comic frames illustrating process steps,* O *is a set of depicted process-relevant objects or actors,* A *is a set of annotated actions or events, and* L *is a layout function that arranges* F *in a coherent temporal or logical order. Each frame in* F *may integrate multimodal elements (e.g., images, text, symbols).*

Definition 2 ***Multimodal Conceptual Modeling (MMCM) Framework:*** *Let* C *be a set of process models, where each* $c \in C$ *is defined as a tuple* $c = (E, R)$*, with* E *being a set of concept elements and* R *a set of relationships between elements in* E*. Let, furthermore,* M *be a set of multimodal evidences, where each* $m \in M$ *is a tuple* $m = (T, V)$*, with* T *being the type of modality (e.g., image, video, text) and* V *the actual content. We define a cross-modal retrieval function* $f : E \to \mathscr{P}(M)$*, linking each concept element to a set of multimodal evidences. Human-to-model communication is represented by the function* $h : M \to E$*, mapping evidence to a corresponding concept element, while model-to-human communication is represented by the function* $g : E \to \mathscr{P}(M)$*, mapping concept elements to sets of multimodal evidences. The framework* $\mathscr{F}$ *for MMCM is thus* $\mathscr{F} = (C, M, f, h, g)$*, enabling bidirectional multimodal communication, allowing modelers to describe concepts using multimodal evidences and models to provide reference examples for concept elements.*

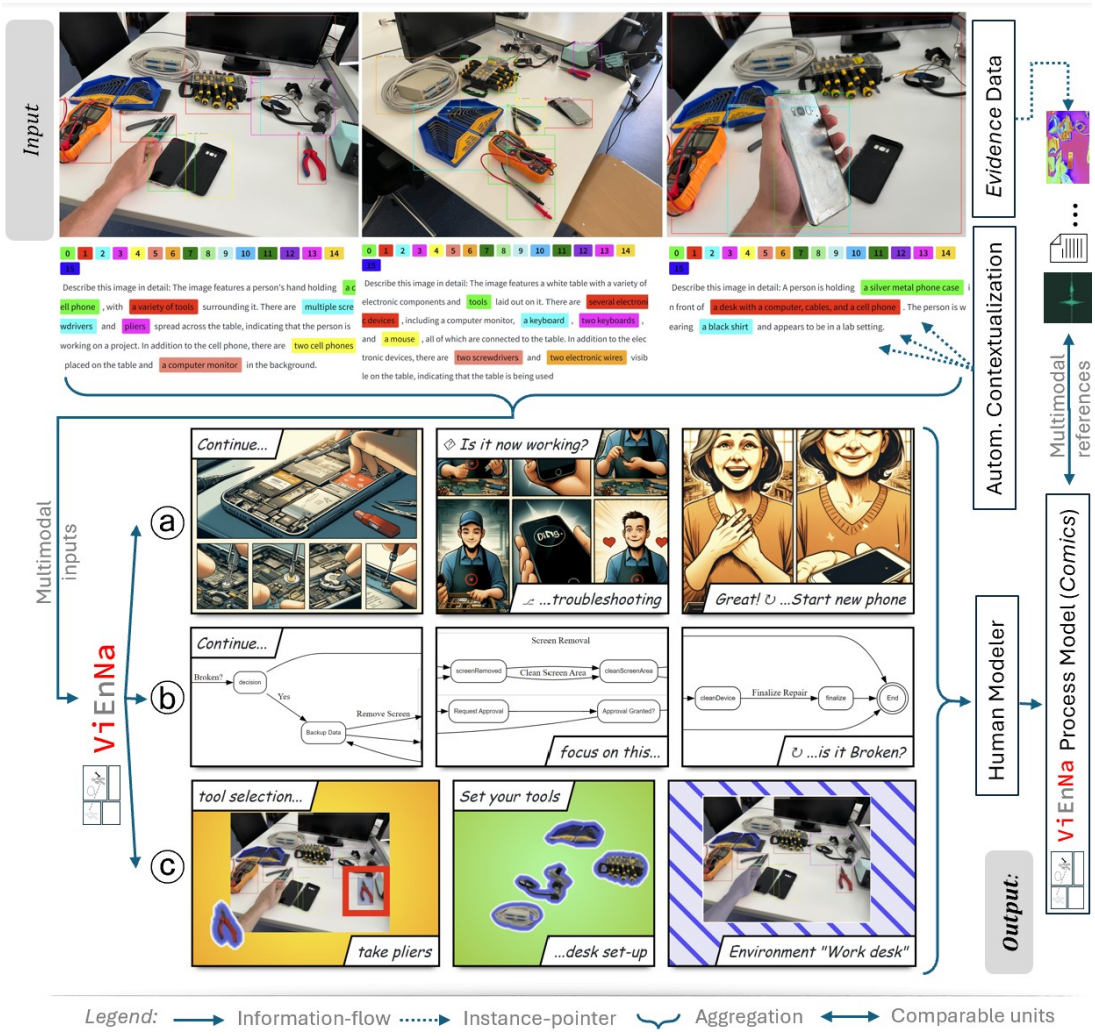


Figure 1: ViEnNa framework illustration. (a) Pixel-level generated panels using generative AI (b) Conceptual model navigation (c) Panels modeled with semantic retrieval. (Example-domain labels are intentionally left unreadable)

We developed ViEnNa as a concrete instantiation of a **process comic** (Def. 1), where the sequence F is generated using AI-based image synthesis conditioned on object-centric event data, the objects O and actions A are extracted from parsed OCEL logs, and the layout function L arranges frames according to temporal and logical process constraints. Furthermore, ViEnNa satisfies the properties of a **Multimodal Conceptual Modeling (MMCM)** framework (Def. 2) by integrating multimodal evidences (e.g., text, visuals, and audio) into each panel.. To establish ViEnNa as an MMCM framework, we developed a system that systematically integrates a suite of AI models, and methods, as detailed in Appendix A.

Intended Use. **(Ways of using)** A modeler can create ViEnNa process models by solely employing a pixel-level generation of comic panels as imagery (Fig. 1a). Here, a modeler can abstract business processes and model complex concepts, including emotions and sound effects, to represent business situations, conditions, or operations, in an engaging way. The second way of using ViEnNa is illustrated in Fig. 1b, where a modeler can depict the storyline behind the reading/interpretation of diagrammatic process models. This approach allows a modeler to wrap other process models into a narrative with an

orchestrated flow of focuses, guiding readers through the concept elements and adding another layer of abstraction to the models. Another way of using ViEnNa process models is shown in Fig. 1c. In this approach, a modeler can discuss concepts at a highly abstract level while the framework understands the semantics behind those concept descriptors and retrieves visual representations from a pool of evidence data materials. Therefore, if a modeler provides videos, audio, images, or a free-form textual file about the concepts being modeled, the framework can comprehend this raw unstructured data and assist in modeling by offering links between the modeled process elements and their actual representations.

While these three examples portray non-overlapping use case scenarios, we designed ViEnNa to facilitate all these approaches as a unified modeling practice. This allows elements of conventional process models (like BPMN) to be linked to the evidence data of a process while modeling additional aspects that are now enabled, ultimately providing a narration-driven scenario modeling. **(Automation)** ViEnNa supports *semi-automated modeling*, where AI assists in interpreting and generating visual narratives from multimodal inputs, while the modeler remains in control of the narrative and structural logic. **(Domains and Users)** The framework is *domain-agnostic* but is particularly effective in domains with rich multimodal evidence (e.g., healthcare, manufacturing, education), and targets a broad range of users—including process modeling experts, non-experts, and AI agents—for both model creation and interpretation tasks. **(Usage by AI)** AI agents can exchange *tremendous amounts of ViEnNa comics per second* among themselves, using them as interpretable, structured representations of knowledge and reasoning steps—ensuring that their collaboration remains transparent and understandable to humans, rather than becoming a black-box exchange, which is a typical scenario Bork et al. (2023); Fill and Muff (2023).

3.1 Synthesizing Comic Image

For image synthesizing, we use the principles of denoising diffusion probabilistic models (Ramesh et al. 2021) for image generation, and *Dall-E* (OpenAI 2024; Ramesh et al. 2021). Starting from Gaussian noise, the process involves learning a series of denoising autoencoders to iteratively refine an image (as illustrated in Fig. 2). Let x_0 be the original image and x_T be the Gaussian noise. The forward diffusion process adds noise to the image in T steps, defined as $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})$, where α_t is a noise schedule. The reverse process, parameterized by a neural network ε_θ, predicts the noise component at each step, given by $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$, where $\mu_\theta(x_t,t)$ and $\Sigma_\theta(x_t,t)$ are functions learned during training. The loss function for training the network is the variational lower bound (VLB) of the data likelihood, expressed as

$$L = \mathbb{E}_q \left[\sum_{t=1}^{T} D_{KL}(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1) \right].$$

This training minimizes the difference between the predicted noise and the actual noise added during the forward process, allowing the model to generate high-quality images by effectively reversing the diffusion process.

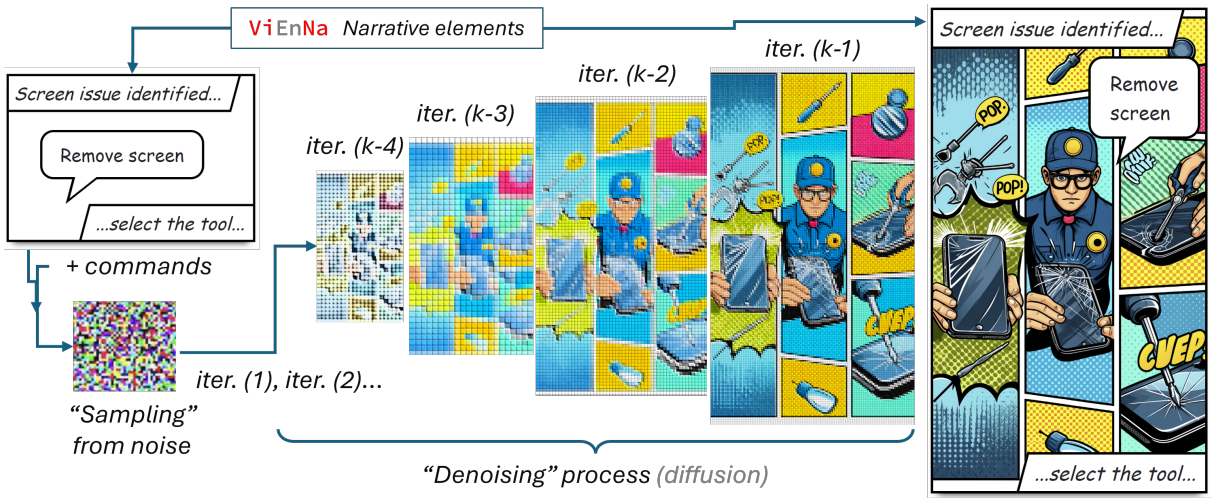


Figure 2: Panel generation (pixel-level) using generative diffusion model.

3.2 Matching Real-World Processes and Event Data

One of the major challenges our framework faces is integrating multimodal evidence data with high-level event logs, particularly object-centric event logs, into a unified process model representation. ViEnNa aim to enable interpretation of high-level event logs based on setting a specific objective through natural language in semantic retrieval and image synthesizing.

Extracting from OCEL 2.0. At first, we focus on extraction of concepts from OCEL 2.0, in particular, *universes of strings* U_Σ, from which we derive various pairwise disjoint universes, as defined by Berti et al. (2024a). These include $U_{ev} \subseteq U_\Sigma$ for events, $U_{etype} \subseteq U_\Sigma$ for event types (activities), $U_{obj} \subseteq U_\Sigma$ for objects, $U_{otype} \subseteq U_\Sigma$ for object types, $U_{attr} \subseteq U_\Sigma$ for attribute names, $U_{val} \subseteq U_\Sigma$ for attribute values, U_{time} for timestamps with $0 \in U_{time}$ as the smallest element and $\infty \in U_{time}$ as the largest, and $U_{qual} \subseteq U_\Sigma$ for qualifiers. Secondly, we use the structures provided in OCEL 2.0 specification to instantiate a tuple $L = (E, O, EA, OA, evtype, time, objtype, eatype, oatype, eaval, oaval, E2O, O2O)$, where $E \subseteq U_{ev}$ is the set of events, $O \subseteq U_{obj}$ is the set of objects, $evtype : E \to U_{etype}$ assigns types to events, $time : E \to U_{time}$ assigns timestamps to events, $EA \subseteq U_{attr}$ is the set of event attributes, $eatype : EA \to U_{etype}$ assigns event types to event attributes, $eaval : (E \times EA) \to U_{val}$ assigns values to event attributes, $objtype : O \to U_{otype}$ assigns types to objects, $OA \subseteq U_{attr}$ is the set of object attributes, $oatype : OA \to U_{otype}$ assigns object types to object attributes, $oaval : (O \times OA \times U_{time}) \to U_{val}$ assigns values to object attributes, $E2O \subseteq E \times U_{qual} \times O$ are the qualified event-to-object relations, and $O2O \subseteq O \times U_{qual} \times O$ are the qualified object-to-object relations.

For instance, consider a manufacturing enterprise where different sensors, logs, and user inputs provide multimodal data. We map these inputs into our defined structure as follows: For any event $e \in E$ and event attribute $ea \in U_{attr}$, we map $eaval_{ea}(e) = eaval(e, ea)$ if $(e, ea) \in \mathrm{dom}(eaval)$; otherwise, it is $\perp$. For example, an event could be the completion of a production task, with attributes like machine ID, operator ID, and timestamp. Similarly, for any object $o \in O$, object attribute $oa \in U_{attr}$, and time $t \in U_{time}$, we map $oaval^t_{oa}(o) = oaval(o, oa, t')$ if there exists $t' \in U_{time}$ such that $t' \leq t$ and no earlier t''

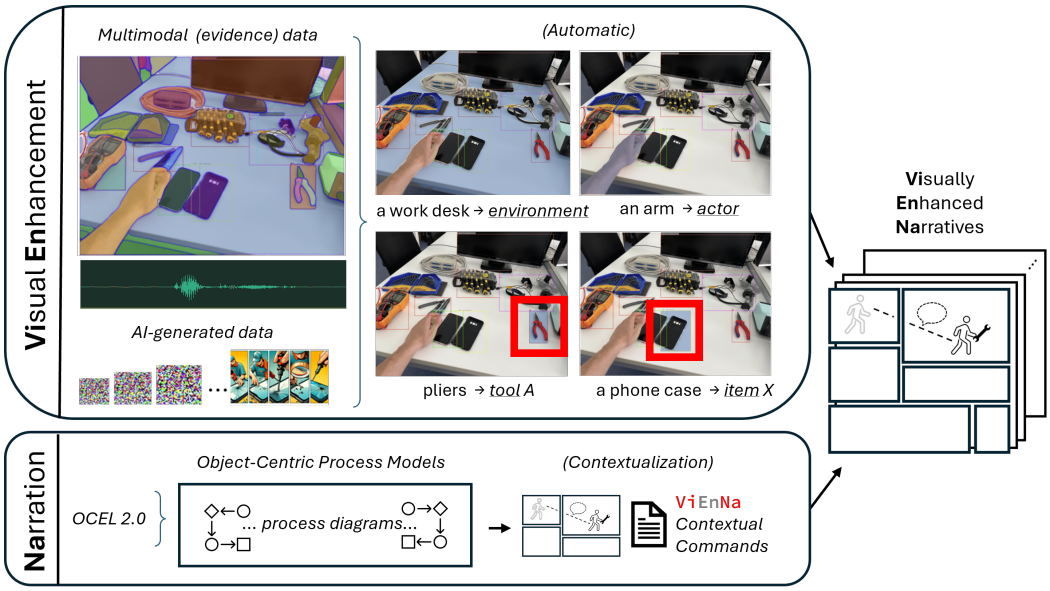


Figure 3: Overview of the implemented solution. *(upper part)* The AI can analyze multimodal data to extract relevant information. *(lower part)* Human modeler is structuring the flow of information, similar to comic strips.

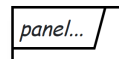
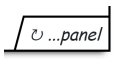
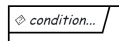
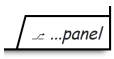
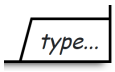
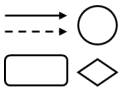
exists that meets the same criteria. If no such t' exists, then $oaval^{t}_{oa}(o) = \perp$. For example, an object could be a specific machine part with attributes like wear level, maintenance status, and temperature readings over time.

Blending OCEL 2.0 and Multimodal Embeddings. The blending of OCEL 2.0 and multimodal embeddings is achieved by setting a *task-centric objective*. This objective organizes the relationships and dynamics between multimodal and event data through human intervention via natural language. We prompt the LLM component of our framework with business-context questions, such as *Who*, *What*, *Why*, *When*, *Where*, *How*, and *How Much*, to analyze both multimodal data and OCEL. This is achieved through two primary methods: (1) retrieving references directly from the multimodal streams, and (2) generating abstractions of these streams through visual or textual synthesis. This way we provide detailed and context-rich answers. To combat hallucinations and false information, we employ principles from retrieval-augmented generation (Lewis et al. 2021) to improve the accuracy and reliability of the generated outputs. This approach combines the strengths of retrieval-based systems and generative models, using the former to ground responses in factual, retrievable data. This approach reduce the risk of generating plausible-sounding but incorrect information, a common issue in purely generative models (Huang et al. 2023).

4 Derivation of the ViEnNa Process Modeling Language

The ViEnNa process modeling language is designed to be used either as a standalone visual process modeling language or as a meta-language to encapsulate other visual process modeling languages into a narrative-driven process model. Although our language can be applied across various business sectors, we specifically tailor it as a domain-specific modeling language, that is particularly focused on processes

Table 1: Narrative graphical notations

Notation	Concept	Description
	*Declaration*	Used to instantiate a new panel within the narrative, introducing a specific scene. This element renders in the top left corner of a panel and has a structure: <panel_name><three_dots>.
	*Unconditional Jump*	Denotes a direct jump or transition in the narrative to another panel, occurring without any conditions. This element renders in the bottom right corner of a panel and has a structure: <jump_symbol><three_dots><destination_panel_name>.
	*Condition*	Represents the instantiation of a panel that is conditional; the narrative within this panel will proceed only if the condition is true. This element renders in the top left corner of a panel and has a structure: <condition_symbol><panel_name><three_dots>.
	*Conditional Jump*	Indicates a jump to another panel that is contingent on a specific condition being met; if the condition is false, the narrative follows the provided alternative path. This element renders in the bottom right corner of a panel and has a structure: <conditional_jump_symbol><three_dots><destination_panel_name>.
	*Typing*	Specifies the type or nature of elements within the narrative, categorizing or clarifying the roles and characteristics of narrative components. This element renders in the bottom right corner of a panel and has a structure: <type_name><three_dots>.
	*Bubble*	Used to display dialogue or thoughts of characters, typically in comics to show what a character is saying or thinking in a visually direct manner. This element renders anywhere inside a panel and has a structure: <bubble_text>.
	*Shapes, Arrows, Labels and Other*	Utilized for representing various process models to be incorporated into ViEnNa, aiding in the visualization and connection of different components and processes within the narrative. These elements can render anywhere inside a panel.

rich in narrative-intensive activities and blind spots introduced by manual work, as described by Kratsch et al. (2022). It is ideally suited for modeling business operations that involve extensive physical and manual tasks with limited human-computer interactions, especially in fields like surgery, construction, and heavy industries. Several methodologies for developing domain-specific languages have been proposed, (Frank 2013; Jannaber et al. 2017; Karsai et al. 2014). We adopt macro-process from Frank (2013), which outlines seven detailed phases (clarification of scope and purpose, analysis of generic requirements, analysis of specific requirements, language specification, design of graphical notation, development of modeling tool, and lastly, evaluation and refinement). We provide a clarification of the scope and purpose, as well as an analysis of both generic and specific requirements, in Appendix B.

4.1 Language Specification and Notation Design

ViEnNa process models are composed of panels arranged in a sequence that is read from left to right, starting from the top row and progressing to the bottom. Within each panel, a multi-language process model, such as a process model, can be displayed, with various parts zoomed in and focused to highlight specific details. The language specification for the ViEnNa process modeling language includes two pivotal components: *Narrative Graphical Notations* and *Visual Enhancements*, each tailored to enhance the storytelling and visual representation of multimodal reference data within a modeling framework, respectively (see Fig. 3).

The Narrative Graphical Notations (as provided in Table 1) are designed to clarify and engage through visual storytelling conventions akin to those found in comics. The narrative begins with the *declaration of a panel*, used to instantiate new scenes within the model. A panel can be referred to by its unique name. To enable modeling concurrency and eliminate the limitation of modeling each state individually, we

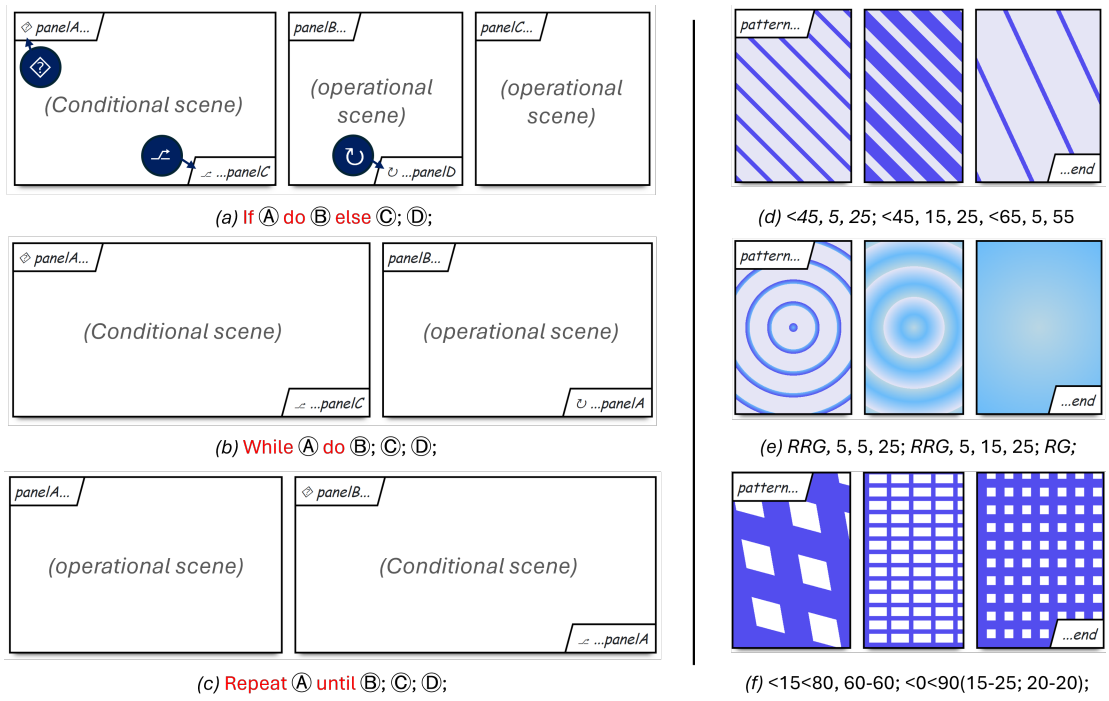
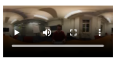


Figure 4: ViEnNa example elements. *Subfigures (a) to (c)* - branching, while and repeat-until control flows. Panels are ordered A, B, C, D (implicit if not visible); *Subfigures (d) to (f)* - pattern coloring in format: (d) and (f) repeating-linear-gradient - angles, positions (e) repeating-radial-gradient (RRG), radial-gradient (RG).

provide conditions and jumps. *Unconditional jumps* allow for direct transitions between panels without any condition, facilitating straightforward narrative progression or flashbacks. Conditional elements are introduced through labeled panels, where the scene within a panel describes a situation. Regular reading flow proceeds only if the condition is true, adding decision-making layers to the flow. A *conditional panel* is followed by a *conditional jump* element that prevents continuation to the next panel depending on the specific condition described within the panel. If the condition is met, an alternative path to a named panel is provided. Different flow structures are illustrated in Fig. 4. *Typing* within the narrative, categorizes and clarifies roles and characteristics of narrative components, and can be used to label or group panels. Typing can be performed either through a text-like graphical element from Fig. 1 or through pattern fill-ups as illustrated in Fig. 4. A *pattern* can be used to symbolize a specific environment where the scene is happening (*e.g., in the office*) or a specific situation (*e.g., a customer enters the shop*). A *legend of used patterns* can be provided. Dialogue or thoughts of characters are prominently displayed through a bubble notation, enhancing narrative interaction. Additionally, various *shapes, arrows, and labels* are utilized to guide the viewer's attention, illustrate connections, and provide additional information, but also for drawing other process models as our platform is providing multi-language modeling support.

Visual Enhancements (as provided in Table 2) focus on visually representing multimodal evidence references to enrich the narrative within a panel. This involves embedding multimedia elements like images, videos, and audio clips directly into the narrative panels, setting scenes, or providing auditory context. *Semantic Retrievals* facilitate cross-referencing between different modalities, ensuring that textual descriptions can be linked to corresponding visual, auditory, or text (file) data. It also includes

Table 2: Notations for visual enhancements with graphical examples

Example	Concept	Description
	*Semantic Image Retrieval*	Involves retrieving images based on natural-language descriptions or other modality inputs, allowing users to find relevant images by describing them in words or using other media as prompts. Syntax: <visualize_image:><concept_descriptor>.
	*Semantic Video Retrieval*	Enables the retrieval of video content using natural-language queries or other types of media, making it possible to locate specific video segments by describing the content or providing related media as input. Syntax: <visualize_video:><concept_descriptor>.
	Semantic Audio Retrieval	Allows users to search for evidence audio clips by using natural-language descriptions or other modalities, enabling efficient access to relevant audio content based on descriptive queries or related media. Syntax: <visualize_audio:><concept_descriptor>.
	*Semantic Textual Retrieval*	Involves retrieving textual information such as documents, tables, or HTML content using natural-language queries or other modality inputs, facilitating the search for specific text-based information through descriptive prompts or related media. Syntax: <visualize_file:><concept_descriptor>.
	*Semantic Pattern Retrieval*	Allows for the retrieval of patterns, such as textures or designs, using natural-language descriptions or other modality inputs, enabling users to find relevant patterns by describing them or using related media. This is cashed from a key-value store. Syntax: <pattern:><concept_descriptor>.
	*Panel Generation*	Retrieves automatic creation (generation) of a visual panel based on natural-language descriptions or other modalities set-up on the scene, allowing users to generate illustrative content by providing descriptive prompts or related media inputs. Syntax: <design:>[<panel_id>|<text_prompt>].

Panel Generation that uses generative AI (as described in Sec. 3.1) to create visual panels from natural-language descriptions with the support of other media inputs. When utilizing panel generation, various customizable options are available to suit specific needs. To name few, *perspectives* such as trainee, client, or decision-maker can be selected to align with the target audience and narrative. *Style choices* range from kid-friendly with vibrant colors and simple illustrations to professional with sleek, polished graphics, and thematic styles tailored to specific business contexts like technology, finance, or healthcare. Additionally, the *level of detail* can be adjusted from minimalistic and straightforward to highly detailed and intricate. These customizable options enable the creation of engaging and relevant content that effectively communicates the intended message.

4.2 Development

The visualization of ViEnNa models is done using a *JavaScript* for frontend operations, ensuring cross-platform HTML rendering. We integrated Graphviz (AT&T 2023) to facilitate the drawing of process models. The final output of our models is generated in *HTML* and *CSS*, offering a standard web-based interface. A custom parser for the ViEnNa meta-language was also designed, allowing for efficient interpretation and execution of the language's syntax and commands for visual rendering. We publish our prototype open source on our GitHub page[1].

5 Evaluation

To empirically validate our contributions, we conducted a comprehensive evaluation covering both *perceived appeal* (Sec. 5.1) and *comprehension accuracy* (Sec. 5.2) of ViEnNa-generated process models. This section presents our hypotheses, outlines the study design (Sec. 5.3), and presents key findings

[1]Implementation and evaluation details are available at: `https://anonymous.4open.science/r/viennacomics/`

based on results from multiple analytical perspectives (Sec. 5.4), further discussed in Sec. 6.

Hypotheses. Our RO be dissected into tasks of *design a modeling framework for comic-like representations of processes that enhances the perceived appeal (APP) of process models—without sacrificing, and ideally improving, comprehension accuracy (ACC)—when compared to existing approaches.* To empirically address the RO we reformulate it into five directional hypotheses, grouped by the two target constructs:

Evaluation Hypotheses for ViEnNa

H_{A1} (Appeal—overall) The weighted perceived-appeal score (*APP*) of *ViEnNa* models is **higher** than that of (i) conventional (i.e. BPMN) models and (ii) alternative non-conventional (comic-like) models not using ViEnNa.

H_{A2} (Appeal—sub-dimensions) ViEnNa outperforms both baselines on each ICE-T (Wall et al. 2019) dimension: Insight (I), Confidence (C), Essence (E), and Time Efficiency (T).

H_{A3} (Appeal—think-aloud visual attention) Readers of ViEnNa devote a **higher proportion of fixations** to semantically relevant regions and complete comprehension tasks with **shorter average dwell time**.

H_{C1} (Comprehension—overall) The weighted comprehension-accuracy score (*ACC*) of ViEnNa is **comparable** than that of conventional models (non-inferiority margin $\delta = -5\%$).

H_{C2} (Comprehension—Beyond Human-Only Use) AI agents based on large language models achieve a higher *PM-LLM-Benchmark* (Berti et al. 2024b) score when augmented with ViEnNa rather than existing conventional and other non-conventional process models.

5.1 Evaluating Perceived Appeal of ViEnNa Models (APP)

The perceived appeal of each modeling approach is evaluated through a weighted combination of three complementary methods: **(APP-1) Heuristic-Based Evaluation (weighted 40%)** employed the ICE-T framework, widely adopted in both the process mining and visual analytics communities (Ciccio et al. 2024), which assesses four key dimensions: Insight Generation (I), Confidence and Trust (C), Essence Conveyance (E), and Time Efficiency (T); **(APP-2) Attention Analysis (weighted 30%)** involved recording participants' gaze to evaluate how effectively information comics, such as ViEnNa models, convey process information. This analysis followed principles from an eye-tracking study on informative comics (Bucher and Boy 2018), using attention heatmaps as the primary metric; Lastly, **(APP-3) Think-Aloud Walkthroughs (weighted 30%)** were conducted to qualitatively assess model comprehension and perceived utility, following established think-aloud protocol guidelines (Haisjackl et al. 2018).

5.2 Evaluating Comprehension Accuracy of ViEnNa Models (ACC)

Comprehension accuracy is evaluated by integrating several process model and narrative evaluation frameworks using the following weight distribution: **(ACC-1) Process Model Comprehension Framework (weighted 40%)** adopted the schema proposed by Winter et al. (2023), which systematically quantifies the comprehension perspectives of both modelers and readers; **(ACC-2) Event Comprehension Theory (weighted 25%)** applied cognitive principles from the Scene Perception and Event Comprehen-

sion Theory (SPECT) to assess the narrative and visual integration from the viewpoint of event perception (Loschky et al. 2018); **(ACC-3) Usage of ViEnNa by AI Agents (weighted 15%)** evaluated whether large language models (LLMs) benefit from using ViEnNa models in process mining tasks, as defined by the PM-LLM-Benchmark (Berti et al. 2024b); **(ACC-4) Multimodal Discourse-Based Comic Analysis (weighted 10%)** assessed the visual discourse structure of comics using the analytical units proposed by Bateman and Wildfeuer (2014a); Lastly, we use **(ACC-5) Narrative Analysis (weighted 10%)** to assess the narrative quality of the models following the approach by Esin (2011).

5.3 Study Design

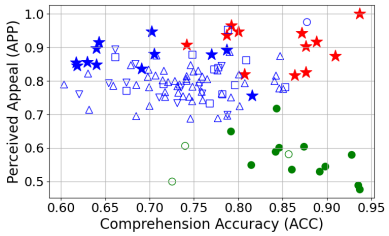
Study Participants. Our participant pool consisted of three groups: 10 PhD holders or candidates in process modeling (*BPM experts*), 12 PhD students from AI-related fields with no background in business process modeling (*AI experts*), and 25 students representing end-users of the modeled processes, such as process trainees (*Non-Experts*). We also developed a platform for gaze tracking—with saccade, and fixation logging—and used it to analyze participants' attention while they read ViEnNa models as well as equivalent conventional and non-conventional process models (as detailed in Appendix C). AI experts were additionally tasked with generating AI-produced comic-like visuals based on textual process descriptions, without using the ViEnNa framework. As for AI agents, we employ four instances of *mistral-small3.1* models through Ollama (2024).

Evaluation Data. **(Domains)** The evaluation employed OCELs from the IT Asset Management (ITAM) domain (Fehrer et al. 2024), comprising 121 process instances, 6 process types, 12 participants (2 IT staff and 10 clients), 6 rooms, over 40 objects (classified into 6 classes), and 100+ events. To ensure cross-domain generalizability, we included two additional datasets: (i) a smaller OCEL on DNA collection from Gavric et al. (2024a), comprising 6 instances, 3 processes, 10 objects, and 12 events; and (ii) a custom LEGO OCEL representing the assembly process of ICPM 2024's LEGO figure of a process mining engineer, containing three instances of a single process, in total of 15 events. **(Novelty)** The LEGO dataset was specifically included to assess how AI models perform on data that have definitely not been seen during training. All OCELs used in the study are accompanied by corresponding videos of the process execution. **(Conventional and Non-Conventional Models)** For each OCEL, we created conventional process models using the Berti and Aalst (2022), and ViEnNa models using our proposed approach. To generate corresponding non-conventional process models, we followed two strategies. First, 15 AI experts were tasked with using AI tools to create comic-like visuals based on textual process descriptions—without employing the ViEnNa framework—resulting in 64 non-conventional models. To ensure quality and relevance, we involved AI experts in crafting these prompts and selecting representative outputs, which allowed us to isolate the added value of ViEnNa's domain-specific modeling features, beyond what generic AI generation alone can offer. **(Bias in Data)** AI experts advised us to include gender-neutral versions of models in our evaluation to examine whether user reactions differ in the presence or absence of gender cues, thereby assessing potential bias in interpretation and engagement. **(Exploratory Experiments)** To explore potential directions for future work, we also introduced an experimental feature—*ViEnNa-Motion*—which replaces static comic panels with short, looping video

clips (up to 5 seconds) to convey dynamic aspects of process activities. This prototype aims to evaluate whether motion-enhanced representations can further improve user comprehension and engagement. We use Brooks et al. (2024) for the video generation, where we provide an image of the ViEnNa panel as an input, followed by a simple prompt that explains the intended use case. As another line of experimentation, we create fully automated variant of ViEnNa—*Auto-ViEnNa*—whose scene-generation prompts are drawn from our curated library of domain-agnostic templates co-crafted with BPM and AI experts. **(Selecting the Baseline)** Second, BPM experts created SAP Scenes following the guide (SAP 2019), resulting in an additional 12 non-conventional models. We select SAP Scenes as the non-conventional modeling approach baseline among other existing non-conventional modeling approaches, because of their established role in professional and training contexts as a tangible, narrative-rich modeling approach that enable hands-on engagement and foster team collaboration through (physical) manipulation of symbolic elements. **(Mitigating Creator Bias)** All manually constructed ViEnNa models are authored by the same individual to eliminate creator-specific bias during evaluation. In contrast, participant-generated non-ViEnNa and non-conventional models may sometimes be self-evaluated, but otherwise are assessed exclusively by other creators, while Auto-ViEnNa, ViEnNa-Motion, and conventional models are all generated automatically. **(Overview)** Resulting data collection comprises models in three domains: 12 ViEnNa models of varying sizes (6 small, 4 medium, 2 large) and their Auto-ViEnNa pairs, 15 conventional process models (12 BPMN and three oc-BPMN equivalents of randomly selected BPMNs), and 77 non-conventional process models—including 64 non-ViEnNa comics (12 of which are gender-neutral to serve as a bias-handling benchmark), 12 SAP Scenes, and one ViEnNa-Motion model. Detailed log of all model creations and the resulting models are available in our supplementary material. An example process with example ViEnNa, conventional and non-conventional model is given in Appendix D.

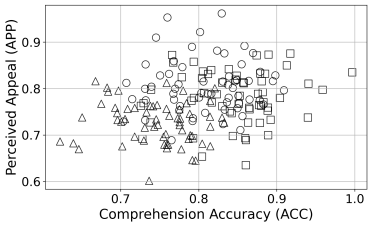
5.4 Evaluation Results

Results are analyzed across three perspectives: **(P-1) ViEnNa vs. Existing Modeling Approaches** – we benchmark ViEnNa (with an inclusion of the Vienna-Motion variation) against two conventional (BPMN, oc-BPMN) and two non-conventional modeling approaches (non-ViEnNa comics with the inclusion of bias-handling cases, SAP Scenes); **(P-2) Model Size and Domain** - comparisons were made across ViEnNa models derived from small (<5 objects, <15 events), medium (<10 objects, <30 events), and large (>10 objects, >30 events) OCELs across three domains (Asset Management, DNA, LEGO); and **(P-3) User Groups** - with contrasted results across reader vs. creator (modeler) roles, expert vs. non-expert participants, and AI agents with and without ViEnNa integration. Figure 5 presents the empirical results of our evaluation, structured along those three analytical perspectives: Fig. 5a compares the perceived appeal and accuracy of ViEnNa (including the ViEnNa-Motion variant) against both conventional (BPMN, oc-BPMN) and non-conventional (SAP Scenes, non-ViEnNa comics) modeling approaches (P-1); Fig. 5b and 5c explore how model size and domain, respectively, affect ViEnNa model performance (P-2); and Fig. 5d–5f analyze differences across user groups including reader vs. creator roles, expert vs. non-expert users, and AI agents with and without ViEnNa integration (P-3).



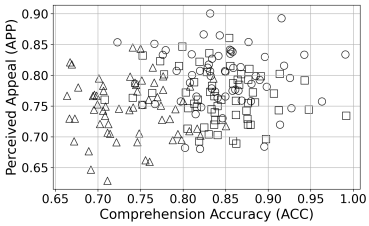
Legend:
★ ViEnNa ★ Auto-ViEnNa
● BPMN ○ oc-BPMN
◯ ViEnNa-Motion ☐ SAP Scenes
▽ Bias tests △ non-ViEnNa Comics

(a) P-1: **ViEnNa (+variants) vs. Existing Modelling Approaches.**



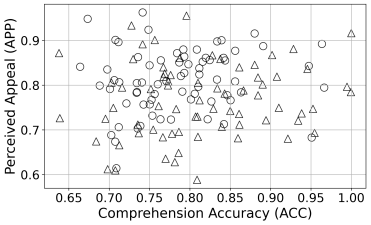
Legend:
☐ Small ◯ Medium △ Large

(b) P-2: **Model Size Experimentation.** *Note:* ViEnNa defines the category mapping function f; $f(\text{OCEL}) \approx f(\text{ViEnNa})$.



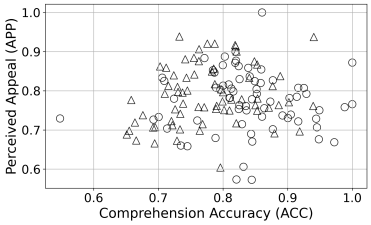
Legend:
☐ Asset Management ◯ DNA
△ *ICPM* LEGO

(c) P-2: **Model Domain Experimentation.** *Note*: as in Fig. 5b.



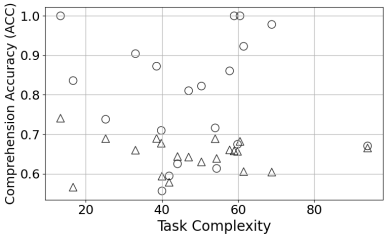
Legend:
◯ Reader △ Creator

(d) P-3: **Reader vs. Creator/Modeler Role Group Experimentation.**



Legend:
◯ Expert △ Non-expert

(e) P-3: **Expert vs. Non-expert User Group Experimentation.**



Legend:
◯ ViEnNa AI △ non-ViEnNa AI

(f) P-3: **AI Agents With vs. Without ViEnNa Integration.**

Figure 5: Results of the evaluation across perspectives (P1-3) from Sec. 5.4.

Evaluation Summary: Revisiting the Research Objective and Hypotheses

H_{A1}: Overall Appeal. Across $N = 116$ participants, *ViEnNa* scored $APP = 0.83$ (SD = 0.05), versus BPMN $\bar{x} = 0.62$ and alt-comics $\bar{x} = 0.58$. One-way ANOVA yielded $F(2, 113) = 105.21$, $p < .001$[a]. Tukey's HSD (FWER = 0.05) showed BPMN–ViEnNa $\Delta = 0.3198$, BPMN–non-ViEnNa Comics $\Delta = 0.2070$, ViEnNa–non-ViEnNa Comics $\Delta = -0.1128$ (all $p < .001$)[b]. **Supported**.

H_{A2}: ICE-T Subscores. MANOVA across the four ICE-T dimensions gave Wilks' $\Lambda = 0.1244$, $F(4, 144) = 66.08$, $p < .001$[c] (Pillai's trace $= 1.2553$, $F(4, 146) = 61.52$, $p < .001$; Hotelling–Lawley $= 3.9877$, $F(4, 85.4) = 71.46$, $p < .001$; Roy's largest root $= 2.9547$, $F(2, 73) = 107.85$, $p < .001$). Follow-ups: Insight, Confidence, Essence ($p \leq 0.002$); Time marginal ($p = 0.07$). **Partially supported**.

H_{A3}: Visual Attention. ViEnNa fixation ratio $M = 0.74$ vs. BPMN $M = 0.63$; $t(35) = 4.9$, $p < .001$, $\delta = 0.54$. Dwell time 12 s vs. 15 s: $t = -3.1$, $p = .004$. **Supported**.

H_{C1}: Comprehension Accuracy. $ACC_{\text{ViEnNa}} = 0.81$ vs. BPMN 0.79. TOST ($\delta = -0.05$) gives $t_1 = 6.5$, $p < .001$; $t_2 = 2.8$, $p = 0.004$. Non-inferiority **supported**, superiority inconclusive ($p = 0.09$).

H_{C2}: Beyond Human-Only Use. Macro-$F_1 = 0.67$ (ViEnNa) vs. 0.55 (BPMN) and 0.52 (text); Wilcoxon $Z = -4.0$, $p < .001$. **Supported**.

Response to the RO. ViEnNa enhances perceived appeal while maintaining or improving comprehension accuracy in both human and AI-assisted settings, fulfilling the research objective.

[a] The F-statistic compares between-group to within-group variance; here, with 3 groups and 116 observations, it indicates a highly significant difference.

[b] Tukey's HSD adjusts for multiple pairwise tests, reporting mean differences Δ and adjusted p-values.

[c] Wilks' Λ is the proportion of multivariate variance not explained by group; lower values imply stronger effects.

Revisiting the Research Objective. To assess the extent to which ViEnNa fulfills its intended role, we revisit the research objective through five targeted hypotheses. These span overall and dimensional measures of user appeal (H_{A1}–H_{A3}) and comprehension (H_{C1}–H_{C2}). The summary in Sect. 5.4 synthesizes key findings to determine whether ViEnNa satisfies criteria central to its design motivation (RO, as discussed in Sect. 5.1 and 5.2).

6 Discussion

Interpretation of Findings. Think-aloud transcripts and comments show that 52% of participants preferred lightly stylized cartoons over realistic renders, with 13% specifically citing a "Ghibli" style as most engaging. Across the four ICE-T dimensions, a MANOVA revealed a strong multivariate effect (Wilks' $\lambda = 0.1244$, $F(4,144) = 66.08$, $p < .001$), with significant follow-ups for Essence, Insight, and Confidence (all $p \leq 0.002$), although Time was only marginally different ($p = 0.07$). Modeling experts rated Essence highest, whereas non-experts prioritized Insight ($M = 0.85$, $SD = 0.06$). Based on comments from creators and readers, creators valued the semi-automated DSML controls for on-the-fly layout tweaks, while readers praised the intuitive panel sequencing. Participants who compared ViEnNa to SAP Scenes reported minimal concern over the absence of haptic feedback in SAP Scenes, scoring low on deprivation ($M = 1.8$, $SD = 0.7$ on a 5-point Likert scale; only 15% rated ≥ 3), indicating they did not substantially miss haptics. In general, participants expressed strong enthusiasm for sound effects and conversational bubbles—onomatopoeic cues like "Whoosh!" evoked heightened excitement ($M = 4.5$, $SD = 0.5$; 78% positive mentions). Moreover, conditional jumps, unconditional jumps, and explicit condition constructs were deemed valuable and were incorporated in approximately 20% of panels (usage rate $= 0.20$, $SD = 0.05$), whereas pattern-based markings appeared in only 5% of panels ($M = 0.05$, $SD = 0.02$).

The LEGO domain proved most challenging—accuracy dipped to 0.75 (vs. 0.81 overall) and dwell time increased by 25%—likely due to its high level of bespoke detail and an uncommon ICPM edition of the LEGO figure; however, this can be readily addressed by adding concise multimodal examples and domain-specific prompt adjustments. Larger models did not negatively impact comprehension accuracy: ViEnNa achieved $ACC = 0.81$ versus 0.79 for BPMN, with non-inferiority supported via two one-sided t-tests ($t_1 = 6.5$, $p < .001$; $t_2 = 2.8$, $p = 0.004$), though superiority remained inconclusive ($p = 0.09$). Perceived appeal (APP) tended to decrease as model size increased, an effect comparably present in both ViEnNa and evaluated alternatives. Non-experts benefited more from using ViEnNa than BPM experts—indicating its accessibility and guidance value—while modelers and readers reported comparable gains.

AI agents collaborating around shared tasks exchanged hundreds of ViEnNa comics as structured communicative artifacts, enabling the transfer of ideas, knowledge, and intentions that remained interpretable to human stakeholders, who overall preferred ViEnNa models. Notably, AI agents demonstrated even greater performance gains from using ViEnNa than human users—achieving a macro-F_1 of 0.67 (vs. 0.55 for BPMN and 0.52 for text; Wilcoxon $Z = -4.0$, $p < .001$).

6.1 Insights from Exploratory Experiments

Automation in ViEnNa Modeling. Automation in ViEnNa Modeling can, in principle, be driven end-to-end: with defined small library of domain-agnostic natural-language prompts (*defaults*), the system parse an OCEL's textual elements into concise scene descriptions and automatically generate both panel layouts and illustrative imagery. In our prototype, these defaults—curated in collaboration with BPM and AI experts explained in Sec. 5.3—serve as a drop-in configuration that requires no additional input from the modeler beyond the initial prompt set. However, our evaluation (as in Fig. 5a) shows that fully automated runs scored substantially lower on both appeal and comprehension: Auto-ViEnNa models achieved a mean APP of 0.58 (SD = 0.07), roughly 30% below semi-automated variants (M = 0.83; t(49) = 8.2, p < .001), and a mean ACC of 0.66 (SD = 0.06), about 18% below semi-automated scores (M = 0.81; t(49) = 6.7, p < .001). This gap likely reflects the diversity of user preferences for visual style, framing, and narrative emphasis, which a one-size-fits-all automation cannot fully anticipate. Therefore, as initially intended, we recommend ViEnNa as a semi-automated framework: while in contrast, the framwork through DSML (see Tab. 1 and Tab. 2) supports fully manual editing but also a fully automated "default" mode. Notably, none of our participants opted for a zero-automation, fully manual workflow—which falls outside ViEnNa's intended usage—and such configurations were omitted from our formal evaluation.

Insights from the Future-Direction Experiment. The exploratory ViEnNa-Motion prototype—where static panels were replaced by 5 s looping video clips—was warmly received. On the APP metric it scored $\bar{x} = 0.96$ ($SD = 0.20$), significantly above the static ViEnNa mean of 0.83 ($t_{49} = 3.2,\ p = .002$). Think-aloud logs supported this observation: 78% of participants spontaneously remarked that the animation improved the storyline's flow or "felt like watching a tutorial". By contrast, the gender-neutral comic variants aimed at bias reduction were less appealing ($\bar{x} = 0.60$); only 18 % of comments were positive, many readers reporting that those figures hampered empathy and role identification, while no significant influence on APP and ACC is observed (Fig. 8). These observations steer our roadmap toward (i) automatically turning ViEnNa storyboards into short explanatory videos, and (ii) building an interactive platform—VR/AR ready—where users can re-contextualise, branch and continue the process narrative in real time while staying tied to the underlying event log.

6.2 Threats to Validity and Challenges

Construct validity was addressed through multi-method triangulation (APP-1…3; ACC-1…5). **Internal validity** threats from learning effects were mitigated by randomized presentation orders and Latin-square counterbalancing. **External validity** is strengthened by three heterogeneous datasets, yet industrial replication remains future work. **Statistical conclusion validity** was guarded via power analysis ($\beta = 0.8$), effect-size reporting, and corrections for multiple comparisons.

Limitations and Trade-offs of ViEnNa as a Methodology. While the evaluation highlights ViEnNa's advantage in perceived appeal and non-inferior comprehension, three methodological caveats limit the generality of these findings. *(i) Dataset and participant scope:* the study relied on relatively synthetic

OCELs and academic trainees/participants; how ViEnNa scales to enterprise-scale logs with sparse evidence or to professional process owners facing audit pressures remains an open question. *(ii) Narrative–formal tension:* ViEnNa's panel-centric syntax relaxes the compositional semantics of BPMN or OC-Petri nets, complicating rigorous verification, simulation, and conformance checking; organisations that need provable properties will incur extra effort to maintain a parallel formal backbone and guard against semantic drift. *(iii) Multimodal Data Conceptualization:* the method relies on high-dimensional embeddings of multimodal data, grounded in curated conceptual models. If these models are modified or become noisy, the ability to accurately interpret processes from raw data may degrade. Consequently, ViEnNa should be positioned as an engaging narrative layer that augments—rather than replaces—traditional, formally executable process-engineering artefacts.

Limitations and Trade-offs of ViEnNa Prototype. Deploying the current prototype of ViEnNa in production environments entails several technical and organisational compromises. First, rendering a storyboard of high-resolution panels while serving real-time cross-modal searches requires either (i) on-premise GPU nodes—substantially more expensive than the CPU-based servers that suffice for most BPM workloads—or (ii) a reliance on commercial *AI-as-a-Service* endpoints, which shifts the capital expenditure to potentially volatile usage fees. Second, because panel content is synthesised from data and generative models, issues of model bias, evidence leakage, and auditability become critical and must be mitigated through redaction pipelines and access controls. A detailed discussion of scalability tests, runtime costs, and the bias-, security-, and privacy-assessment protocol for both deployment paths is provided in Appendix E.

7 Conclusion and Outlook

In conclusion, the development of the ViEnNa framework marks an important step forward in enriching human-model interaction by integrating multimodal evidence and narrative elements into process modeling, thus matching real-world processes and event data. ViEnNa is designed as a process modeling tool, where the responsibility for creating business-domain accurate models lies with the modeler, who supervises the generation process. Empirical results across multiple domains show that ViEnNa boosts perceived appeal, improves (or occasionally sustains) comprehension accuracy, sharpens reader focus, and helps AI agents reason more effectively about processes.

Future Work. Next steps are structured as follows: *(A) Animate ViEnNa storyboards* — automatically convert comic panels into short video clips or full tutorial sequences by adding motion, sound, and voice-over, while preserving event-log traceability; *(B) Enable mixed-reality guidance* — use AR/VR overlays to project ViEnNa models that walk frontline workers through process tasks and allow trainees to "step inside" the process logic; *(C) Control conceptual drift* — maintain alignment between multimodal embeddings and evolving domain ontologies to ensure semantic fidelity across generated media; *(D) Run longitudinal field studies* — deploy ViEnNa in real-world BPM scenarios to assess its sustained impact on collaboration, comprehension, and decision-making quality over time. By extending comics

into immersive audiovisual experiences, tightening semantic control, and validating impact in the field, ViEnNa aspires to turn process knowledge into living, instructive environments for humans and AI alike.

References

van der Aalst WMP, Berti A (2020) Discovering object-centric petri nets. Fundam Informaticae 175:1–40, 10.3233/FI-2020-1946

van der Aalst WMP, Carmona J (eds) (2022) Process Mining Handbook, Lecture Notes in Business Information Processing, vol 448. Springer, 10.1007/978-3-031-08848-3

AT&T (2023) Graphviz - graph visualization software. `https://graphviz.org/`

Bardes A, Garrido Q, Ponce J, Chen X, Rabbat M, LeCun Y, Assran M, Ballas N (2024) Revisiting feature prediction for learning visual representations from video. `2404.08471`

Bateman JA, Wildfeuer J (2014a) Defining units of analysis for the systematic analysis of comics: A discourse-based approach. Studies in Comics 5(2):373–403

Bateman JA, Wildfeuer J (2014b) A multimodal discourse theory of visual narrative. Journal of Pragmatics 74:180–208

Benzin JV, Park G, Rinderle-Ma S (2023) Preventing object-centric discovery of unsound process models for object interactions with loops in collaborative systems: Extended version. ArXiv 10.48550/arXiv.2303.16680

Berti A, Aalst W (2022) Oc-pm: analyzing object-centric event logs and process models. International Journal on Software Tools for Technology Transfer 25:1–17, 10.1007/s10009-022-00668-w

Berti A, Koren I, Adams JN, Park G, Knopp B, Graves N, Rafiei M, Liß L, Unterberg LTG, Zhang Y, Schwanen C, Pegoraro M, van der Aalst WMP (2024a) Ocel (object-centric event log) 2.0 specification. `https://arxiv.org/abs/2403.01975`, `2403.01975`

Berti A, Kourani H, van der Aalst WM (2024b) Pm-llm-benchmark: Evaluating large language models on process mining tasks. In: International conference on process mining, Springer, pp 610–623

Bork D, Ali SJ, Roelens B (2023) Conceptual modeling and artificial intelligence: A systematic mapping study. CoRR abs/2303.06758, 10.48550/ARXIV.2303.06758, 2303.06758

Brooks T, Peebles B, Holmes C, DePue W, Guo Y, Jing L, Schnurr D, Taylor J, Luhman T, Luhman E, Ng C, Wang R, Ramesh A (2024) Video generation models as world simulators https://openai.com/research/video-generation-models-as-world-simulators

Bucher HJ, Boy B (2018) How informative are information comics in science communication? empirical results from an eye-tracking study and knowledge testing. In: Empirical comics research, Routledge, pp 176–196

Charles R, et al. (2024) Master thesis: Generative ai methods to create comic strips

Chen Y, Jhala A (2024) Collaborative comic generation: Integrating visual narrative theories with AI models for enhanced creativity. In: Filippo AD, Pachet F, Presutti V, Steels L (eds) Proceedings of the 3rd workshop on artificial intelligence and creativity co-located with 27th european conference on artificial intelligence (ECAI 2024), santiago de compostela, spain, october 20, 2024, CEUR-WS.org, CEUR Workshop Proceedings, vol 3810, pp 98–111, https://ceur-ws.org/Vol-3810/paper8.pdf

Ciccio CD, Miksch S, Soffer P, Weber B, Meroni G (2024) Human in the (process) mines (dagstuhl seminar 23271). Dagstuhl Reports 13(7):1–33, 10.4230/DagRep.13.7.1, https://drops.dagstuhl.de/entities/document/10.4230/DagRep.13.7.1

Dash S, Lyngaas I, Yin J, Wang X, Egele R, Cong G, Wang F, Balaprakash P (2023) Optimizing distributed training on frontier for large language models. 2312.12705

Ericsson L, Gouk H, Loy CC, Hospedales TM (2022) Self-supervised representation learning: Introduc-

tion, advances, and challenges. IEEE Signal Processing Magazine 39(3):42–62, 10.1109/msp.2021.3134634, 10.1109/MSP.2021.3134634

Esin C (2011) Narrative analysis approaches. Qualitative research methods in psychology: Combining core approaches pp 92–117

Fehrer T, Egger A, Chvirova D, Wittmann J, Wördehoff N, Kratsch W, Röglinger M (2024) Business Processes in IT Asset Management Multimedia Event Log. 10.6084/m9.figshare.25246291

Fettke P, Reisig W (2020) Modelling service-oriented systems and cloud services with heraklit. 2009.14040

Feuerriegel S, Hartmann J, Janiesch C, Zschech P (2024) Generative AI. Bus Inf Syst Eng 66(1):111–126, 10.1007/S12599-023-00834-7, https://doi.org/10.1007/s12599-023-00834-7

Fill H (2024) Spatial conceptual modeling: Anchoring knowledge in the real world. CoRR abs/2407.17259, 10.48550/ARXIV.2407.17259, https://doi.org/10.48550/arXiv.2407.17259, 2407.17259

Fill H, Muff F (2023) Visualization in the era of artificial intelligence: Experiments for creating structural visualizations by prompting large language models. CoRR abs/2305.03380, 10.48550/ARXIV.2305.03380, https://doi.org/10.48550/arXiv.2305.03380, 2305.03380

Fill H, Fettke P, Köpke J (2023) Conceptual modeling and large language models: Impressions from first experiments with chatgpt. Enterp Model Inf Syst Archit Int J Concept Model 18:3, 10.18417/EMISA.18.3, https://doi.org/10.18417/emisa.18.3

Frank U (2013) Domain-Specific Modeling Languages: Requirements Analysis and Design Guidelines, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 133–157. 10.1007/978-3-642-36654-3_6

Garrido Q, Assran M, Ballas N, Bardes A, Najman L, LeCun Y (2024) Learning and leveraging world models in visual representation learning. 2403.00504

Gavric A, Bork D, Proper HA (2024a) Enriching business process event logs with multimodal evidence.

In: Paja E, Zdravkovic J, Kavakli E, Stirna J (eds) The practice of enterprise modeling - 17th IFIP working conference, poem 2024, stockholm, sweden, december 3-5, 2024, proceedings, Springer, Lecture Notes in Business Information Processing, vol 538, pp 175–191, 10.1007/978-3-031-77908-4_11, https://doi.org/10.1007/978-3-031-77908-4_11

Gavric A, Bork D, Proper HA (2024b) Multimodal process mining. In: 26th international conference on business informatics, CBI 2024, vienna, austria, september 9-13, 2024, IEEE, pp 99–108, 10.1109/CBI62504.2024.00021, https://doi.org/10.1109/CBI62504.2024.00021

Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I (2023) Imagebind: One embedding space to bind them all. In: Cvpr

Goossens A, Smedt JD, Vanthienen J, van der Aalst WMP (2022) Enhancing data-awareness of object-centric event logs. ArXiv 10.48550/arXiv.2212.02858

Grisold T, Seidel S, Heck M, Berente N (2024) Digital surveillance in organizations. Bus Inf Syst Eng 66(3):401–410, 10.1007/S12599-024-00866-7, https://doi.org/10.1007/s12599-024-00866-7

Guérin C, Rigaud C, Bertet K, Revel A (2017) An ontology-based framework for the automated analysis and interpretation of comic books' images. Information sciences 378:109–130

Haisjackl C, Soffer P, Lim SY, Weber B (2018) How do humans inspect bpmn models: an exploratory study. Software & Systems Modeling 17(2):655–673

Heidrich D, Schreiber A, Theis S (2024) Generative artificial intelligence for the visualization of source code as comics. In: International conference on human-computer interaction, Springer, pp 35–49

Herrmann DA, Levinstein BA (2024) Standards for belief representations in llms. 2405.21030

Hofer S, Schwentner H, Safari aOMC (2021) Domain Storytelling: A Collaborative, Visual, and Agile Way to Build Domain-Driven Software. Addison-Wesley Professional, https://books.google.at/books?id=A3yfzgEACAAJ

Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T (2023) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. https://arxiv.org/abs/2311.05232, 2311.05232

Jannaber S, Riehle DM, Delfmann P, Thomas O, Becker J (2017) Designing a framework for the development of domain-specific process modelling languages. In: Maedche A, vom Brocke J, Hevner A (eds) Designing the digital transformation, Springer International Publishing, Cham, pp 39–54

Karsai G, Krahn H, Pinkernell C, Rumpe B, Schindler M, Völkel S (2014) Design guidelines for domain specific languages. arXiv preprint arXiv:14092378 1409.2378

Ketola A, de Rooy R, Haapio H (2024) Comic contracts 2.0–contracts that have (and give) a voice. Design (s) for Law

Kim S, Lee S, Kim K, Oh U (2024) Utilizing a dense video captioning technique for generating image descriptions of comics for people with visual impairments. In: Proceedings of the 29th international conference on intelligent user interfaces, pp 750–760

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R (2023) Segment anything. https://arxiv.org/abs/2304.02643, 2304.02643

Kratsch W, König F, Röglinger M (2022) Shedding light on blind spots – developing a reference architecture to leverage video data for process mining. Decision Support Systems 158:113,794, https://doi.org/10.1016/j.dss.2022.113794, https://www.sciencedirect.com/science/article/pii/S0167923622000653

Kumar JT, Babu H, Srinivasan N (2024) Comic generation using ai–a review. In: International conference on computational intelligence in data science, Springer, pp 156–166

Laubrock J, Dunst A (2020) Computational approaches to comics analysis. Topics in cognitive science 12(1):274–310

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, tau Yih W, Rocktäschel T, Riedel S, Kiela D (2021) Retrieval-augmented generation for knowledge-intensive nlp tasks. https://arxiv.org/abs/2005.11401, 2005.11401

Liu H, Li C, Wu Q, Lee YJ (2023) Visual instruction tuning. 2304.08485

Loschky LC, Hutson JP, Smith ME, Smith TJ, Magliano JP (2018) Viewing static visual narratives through the lens of the scene perception and event comprehension theory (spect). In: Empirical comics research, Routledge, pp 217–238

Lukyanenko R, Bork D, Storey VC, Parsons J, Pastor O (2023) Inclusive conceptual modeling: Diversity, equity, involvement, and belonging in conceptual modeling (short paper). In: Companion proceedings of the 42nd international conference on conceptual modeling: ER forum, 7th scme, project exhibitions, posters and demos, and doctoral consortium co-located with ER 2023, lisbon, portugal, november 06-09, 2023, CEUR-WS.org, CEUR Workshop Proceedings, vol 3618

Malach E (2023) Auto-regressive next-token predictors are universal learners. 2309.06979

Miron ET, Muck C, Karagiannis D (2019) Transforming haptic storyboards into diagrammatic models: The scene2model tool. In: Hawaii international conference on system sciences, https://api.semanticscholar.org/CorpusID:102351683

Muff F, Fill H (2024) Limitations of chatgpt in conceptual modeling: Insights from experiments in metamodeling. In: Giese H, Rosenthal K (eds) Modellierung 2024 - workshop proceedings, potsdam, germany, march 12-15, 2024, Gesellschaft für Informatik e.V., p 8, 10.18420/MODELLIERUNG2024-WS-008, https://doi.org/10.18420/modellierung2024-ws-008

Naur P, Randell B, Committee NS (1969) Software Engineering: Report of a Conference Sponsored by the NATO Science Committee, Garmisch, Germany, 7th to 11th October, 1968. Scientific Affairs Division, NATO, https://books.google.rs/books?id=Uc9QAAAAYAAJ

Nguyen NV, Rigaud C, Burie JC (2019) Comic mtl: optimized multi-task learning for comic book image

analysis. International Journal on Document Analysis and Recognition (IJDAR) 22:265–284

Object Management Group (2013) Business process model and notation (bpmn), version 2.0.2. http://www.omg.org/spec/BPMN, version 2.0.2 contains a minor change to Clause 15.

Ollama (2024) Ollama website. https://www.ollama.com/, accessed: 2024-06-01

OpenAI (2024) Dall·e 3. https://openai.com/index/dall-e-3/, accessed: 2024-06-01

Papoutsaki A, Sangkloy P, Laskey J, Daskalova N, Huang J, Hays J (2016) WebGazer: Scalable webcam eye tracking using user interactions. In: Proceedings of the 25th international joint conference on artificial intelligence (ijcai-16), AAAI, pp 3839–3845

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I (2021) Zero-shot text-to-image generation. 2102.12092

Rebmann A, Rehse JR, van der Aa H (2022) Uncovering object-centric data in classical event logs for the automated transformation from xes to ocel 10.1007/978-3-031-16103-2_25

Rohrer T, Ghahfarokhi AF, Behery MH, Lakemeyer G, van der Aalst WMP (2022) Predictive object-centric process monitoring. ArXiv 10.48550/arXiv.2207.10017

SAP (2019) Scenes concept and building guide. https://apphaus.sap.com/wp-content/uploads/sites/2/2019/07/ScenesConceptAndBuildingGuidepdf-2.pdf, accessed: 2024-06-01

Singh A (2023) A review on objective-driven artificial intelligence. 2308.10135

Tanaka T, Shoji K, Toyama F, Miyamichi J (2007) Layout analysis of tree-structured scene frames in comic images. In: Ijcai, Citeseer, vol 7, pp 2885–2890

Völz A, Muck C, Utz W (2024) Digital twins for haptic design thinking: An innovative prototype. In: 19. internationale tagung wirtschaftsinformatik, Wirtschaftsinformatik 2024 Proceedings, http://eprints.cs.univie.ac.at/8162/

Wall E, Agnihotri M, Matzen L, Divis K, Haass M, Endert A, Stasko J (2019) A heuristic approach to

value-driven evaluation of visualizations. IEEE Transactions on Visualization and Computer Graphics 25(1):491–500, 10.1109/TVCG.2018.2865146

Winter M, Pryss R, Fink M, Reichert M (2023) Towards measuring and quantifying the comprehensibility of process models: the process model comprehension framework. Information Systems and e-Business Management 21(3):723–751

Xiu B, Li G (2023) Diagnosing conformance between object-centric event logs and models. IEEE Access 11:110,837–110,849, 10.1109/ACCESS.2023.3322366

Xu J, Moor M, Leskovec J (2024) Reverse image retrieval cues parametric memory in multimodal llms. 2405.18740

A Multimodal AI Capabilities of ViEnNa

Current AI systems are missing a significant component: the integration of multimodal domain process knowledge (Garrido et al. 2024). Essentially, systems are required, that are capable of learning how the world operates (Herrmann and Levinstein 2024) not solely from text but also from videos and other sensory inputs. Furthermore, these systems must have persistent memory about concepts and processes, a feature absent in current AI technologies (Xu et al. 2024). They must be able to plan actions to fulfill objectives and be controllable and safe, possibly through the specification of guardrail objectives which we can bring through process models - once the model is created, we adhere to it. This encapsulates the concept of objective-driven AI architectures (Singh 2023). Central to the implementation of our solution is the concept of self-supervised learning. Self-supervised learning (Ericsson et al. 2022) involves taking a piece of data, such as text, and transforming or corrupting it in some manner. For instance, in the context of textual data, certain words may be replaced by blank markers. Subsequently, a large neural network is trained to predict the missing words, thereby reconstructing the original input. This methodology underpins the training of large language models, who, once trained, can take a sequence of words and predict the subsequent word. This iterative process continues, with each new prediction becoming

part of the input, a method known as auto-regressive prediction (Malach 2023). These models perform exceptionally well despite their process simplicity, and they are typically trained on vast amounts of data – often around 20 trillion tokens (Dash et al. 2023), where a token is a sub-word unit averaging three-quarters of a word.

To extend the power of self-supervised learning to other data modalities such as images, audio, and more, we use a method to tokenize these modalities similarly to how we tokenize text. At the core of our multimodal process retrieval is the concept of tokenization and embedding. Tokenization involves breaking down each modality into a set of discrete tokens, which are then embedded into a continuous vector space. After obtaining embeddings for each modality, the next step is to align them into a common vector space. This is achieved through a shared encoder network that maps modality-specific embeddings to a unified space. Let $\mathbf{E}_T$, $\mathbf{E}_I$, $\mathbf{E}_A$, etc., be the embeddings for text, image, audio, and other modalities, respectively. The tokenization function $\mathscr{T}$ converts raw data D into tokens $\mathbf{T}$ as $\mathbf{T} = \mathscr{T}(D)$. The embedding function $\mathscr{E}$ maps tokens to embedding vectors $\mathbf{E}$ as $\mathbf{E} = \mathscr{E}(\mathbf{T})$. Finally, the shared encoder function f aligns embeddings into a common vector space as $\mathbf{U} = f(\mathbf{E})$.

We can apply this method in cross-modal retrieval, where a query in one modality (e.g., text) retrieves relevant data in another modality (e.g., images). When a domain modeler queries a concept, we tokenize and embed the query Q (text) and the target data D (images) as $\mathbf{E}_Q = \mathscr{E}_T(\mathscr{T}_T(Q))$ and $\mathbf{E}_D = \mathscr{E}_I(\mathscr{T}_I(D))$ (see Fig. 6). Then, we map embeddings to the common space as $\mathbf{U}_Q = f(\mathbf{E}_Q)$ and $\mathbf{U}_D = f(\mathbf{E}_D)$. We compute similarity (in particular, cosine similarity) in the common space to find the best match:

$$\text{Similarity}(\mathbf{U}_Q, \mathbf{U}_D) = \frac{\mathbf{U}_Q \cdot \mathbf{U}_D}{\|\mathbf{U}_Q\| \|\mathbf{U}_D\|}$$

We envision a process modeling approach built upon the foundations of cross-modal retrieval, linking elements of process models to actual multimodal evidence of these concepts in practice. In such modeling, a modeler can show a picture or a video and use it to model concepts on an abstract level. This represents the human-to-model way of multimodal communication. Conversely, the model-to-human way of

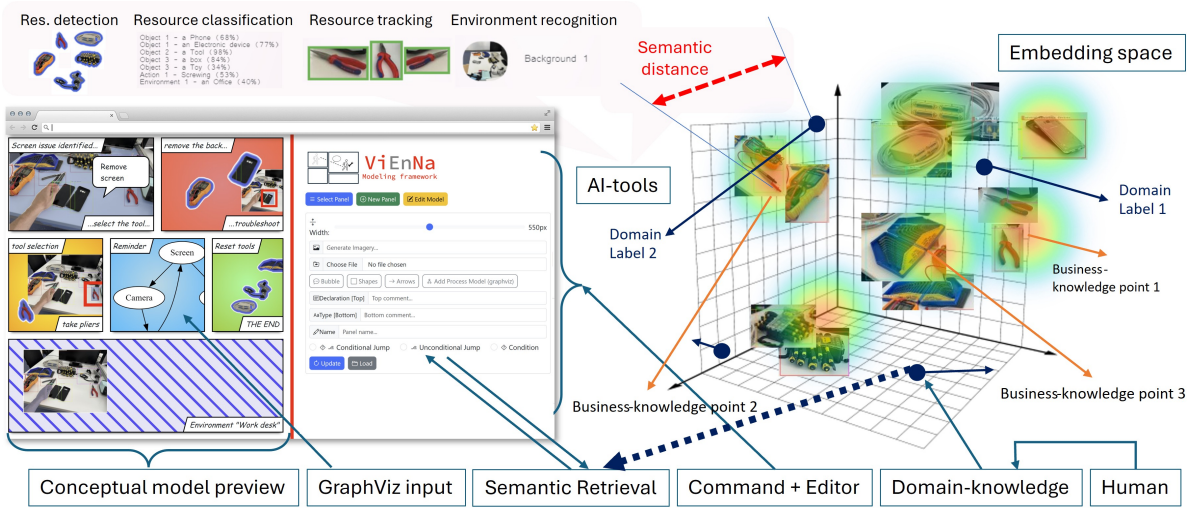


Figure 6: Illustration of our semantic concept retrieval. *(left part)* Implemented modeling tool with a semantic prompt. *(right part) Embedding space* (a.k.a. *Latent space*). (Example-domain labels intentionally left unreadable)

multimodal communication means that a model is capable of showing a reference example modality for a given concept element.

To establish ViEnNa as an MMCM framework in accordance with Def. 2, we developed a system that systematically integrates a suite of AI tools and methods. For the foundational base AI model, we employ Meta's ImageBind (Girdhar et al. 2023). This model is designed to integrate various data modalities into a unified embedding space. ImageBind supports a range of modalities including text, images, audio, depth, thermal imaging, and inertial measurement units (IMUs), by using modalities paired with images (e.g., (I,M) where I represents images and M is another modality) to create a unified embedding using web datasets with image-text pairings covering extensive semantic concepts. For each modality pair (I,M), images I_i and their corresponding modalities M_i are encoded into normalized embeddings q_i and k_i using the functions

$$q_i = \frac{f(I_i)}{\|f(I_i)\|} \quad \text{and} \quad k_i = \frac{g(M_i)}{\|g(M_i)\|}$$

where f and g are deep networks. The embeddings optimize using the InfoNCE loss (Girdhar et al. 2023):

$$L_{T,M} = -\log\left(\frac{\exp(q_i^T k_i/\tau)}{\exp(q_i^T k_i/\tau) + \sum_{j\neq i}\exp(q_i^T k_j/\tau)}\right).$$

Here, τ is a temperature parameter smoothing the softmax distribution, and j denotes negatives or unre-

lated observations. A symmetric loss $L_{T,M} + L_{M,T}$ is used for alignment. For the segmentation model, we use (Kirillov et al. 2023). We use a visual language model, LLaVA2 (Liu et al. 2023) from Ollama (Ollama 2024), to estimate the relationships between extracted resources and their descriptive features.

B Derivation of the Modeling Language

Clarification of Scope and Purpose ViEnNa aims to make process modeling more accessible and useful, especially in narrative-driven scenarios where traditional models fall short. By emphasizing a narrative-driven approach and focusing on visual cues and the sequential flow of events, it helps all stakeholders—regardless of their technical background—understand complex processes more intuitively. This not only improves communication and training but also enhances documentation and compliance for processes that are typically difficult to track.

Analysis of Generic Requirements For the ViEnNa process modeling language, the development framework is structured around three essential General Requirements, corresponding to our three fundamental pillars – narration, deduction, and formation.

GR-1: Narration: *The language must effectively support narrative-driven processes, enhancing storytelling in environments that involve significant physical and manual work.*

GR-2: Deduction *The language needs to include analytical, predictive, and generative tools to aid in decision-making and effectively utilize multimodal evidence, thereby enabling users to extract actionable insights from the modeled scenarios.*

GR-3: Formation: *It should integrate seamlessly with existing major modeling languages and provide accurate representations of diverse system architectures, ensuring it is adaptable and precise for modeling complex systems.*

Analysis of Specific Requirements To translate the broad goals set by the general requirements into actionable steps, we have outlined specific requirements (**SR-1** to **SR-13**). These detailed criteria help

implement the principles of *Narration*, *Deduction*, and *Formation* in the modeling language, ensuring it meets the unique needs of different applications effectively. *GR-1* aims to enhance the narrative capability of the language, making it intuitive and engaging. *The modeling language must be designed to offer visual clarity and intuitiveness, facilitating immediate understanding and ease of use for all users* **(SR-1)**. *It should incorporate features that promote interactivity and sustain user engagement throughout the modeling process* **(SR-2)**. Additionally, *the language needs to support narrative flows centered around characters or key elements, ensuring coherence and engaging storytelling* **(SR-3)**. *The modeling language must be easy to learn, allowing new users to quickly become proficient* **(SR-4)**. *It must effectively represent time, allowing for an accurate depiction of scenarios as they evolve* **(SR-5)**. *The language should accurately represent the dynamics within scenarios, highlighting changes and interactions* **(SR-6)**. *GR-2* focuses on enabling users to analyze data and predict outcomes through the modeling process. *Predictive analytics capabilities must be integrated into the language, providing implicit predictions about enriching narrative scenarios based on provided data* **(SR-7)**. *The language should be capable of incorporating multimodal evidence (text, images, video) to enhance the richness of the narratives and support modeling tasks* **(SR-8)**. *It needs to possess generative capabilities to autonomously produce data and scenarios, enhancing its predictive and analytical utility* **(SR-9)**. *GR-3* ensures the structural adaptability of the language, aligning it with existing frameworks and accurately depicting both, physical and digital, systems architectures. *The new modeling language must integrate with existing models and frameworks* **(SR-10)**. *It must operate effectively across a full spectrum of abstraction, from high-level process overviews to detailed, specific artifacts* **(SR-11)**. *The language must represent both, physical and digital architecture of the systems it models* **(SR-12)**. *It should include a precise digital versioning system to efficiently manage and track changes in models* **(SR-13)**.

C Developed Evaluation Framework

We developed a browser-based experimental platform (shown in Fig. 7) that integrates real-time gaze and time-to-completion tracking, survey prompts, and user response recording.

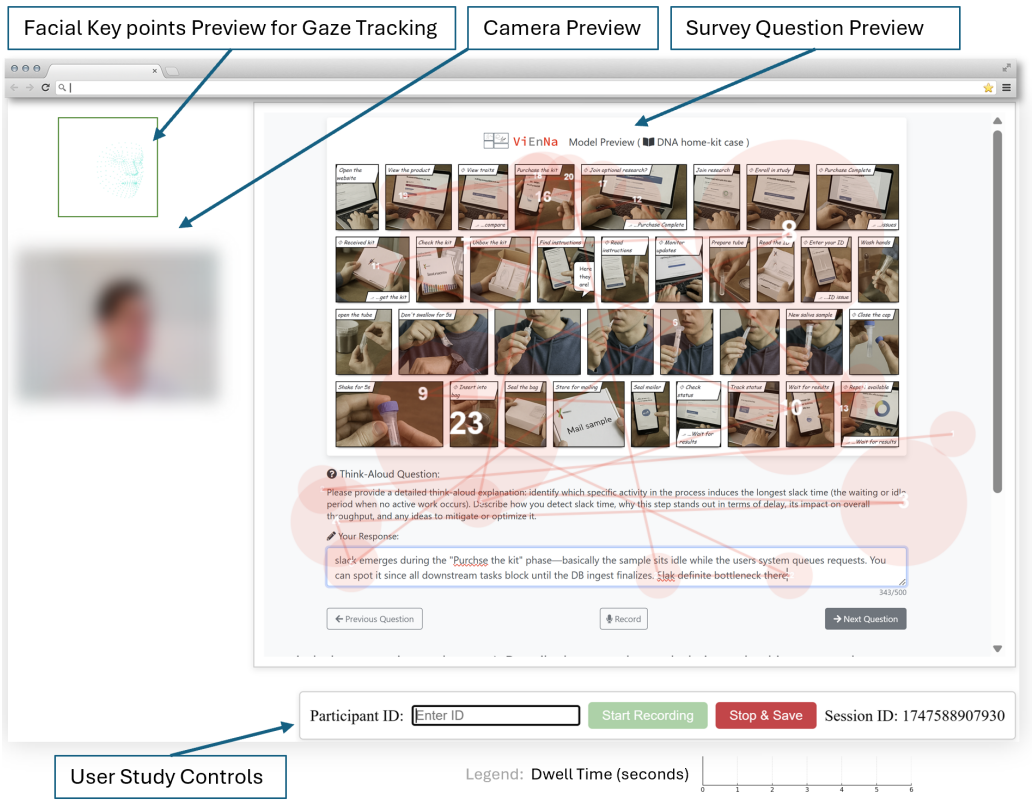


Figure 7: ViEnNa evaluation interface with integrated gaze tracking, response capture, and (post experiment) dwell-time visualization. The numbers in the dwell time circles indicate the order of fixations (smaller numbers denote earlier fixations). Saccades are shown as connecting lines between fixations.

Our system uses *WebGazer.js* (Papoutsaki et al. 2016) to capture participants' eye movements via webcam. A preview of the detected facial key points is shown (top-left) to ensure proper calibration. Simultaneously, the blurred camera preview reassures the participant that the system is functioning and that they remain within the camera's field of view throughout the session.

Visual Stimuli Presentation. At the core of the interface is the *model preview* section, where ViEnNa models are presented inside of the *survey question preview* window (as well as conventional and non-conventional models in other examples). Post-experiment, the survey window is annotated with gaze overlays showing dwell time intensity—red circles—indicating the participant's fixation duration on different activities. Lines connecting circles depict saccades, offering insights into visual scanpaths.

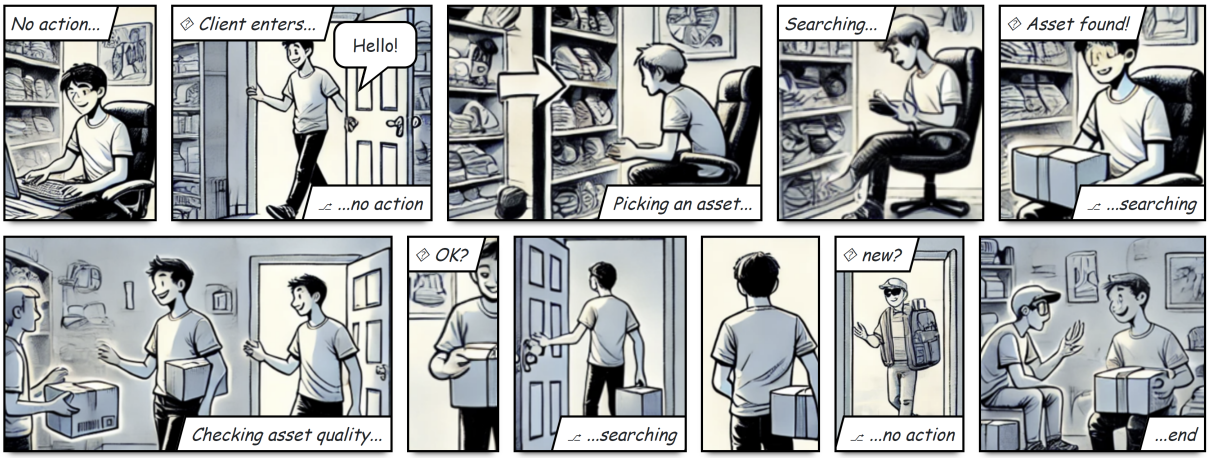
Survey Integration and Think-Aloud Capture. Beneath the model, we integrate a question box that prompts participants to reflect on cognitive aspects of process understanding (e.g., identifying slack time), as described in Sec. 5.

Postprocessing and Metrics. Gaze data is post-processed to extract three key metrics: (1) *Fixation maps*: heatmaps overlaying fixations onto process model regions. (2) *Saccade paths*: sequences and transitions among fixated elements. (3) *Dwell time distributions*: element-wise statistics of visual attention. These allow both quantitative comparison across models (e.g., ViEnNa vs. BPMN) and qualitative insights (e.g., hotspots around narrative transitions). For photorealistic models like SAP Scenes, we display photographs of scenes digitally to preserve compatibility with the gaze tracking.

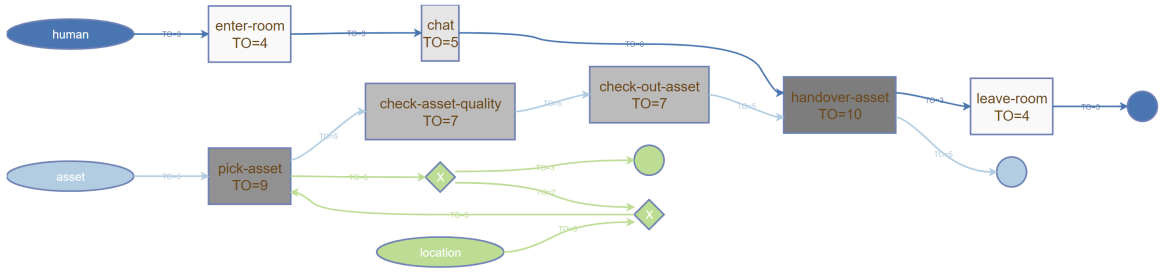
D Example ViEnNa Model

In this section, we present a comparison of multiple representations of the same process derived from multimodal data and OCEL. Figure 8 illustrates three distinct representations of the same process derived from identical multimodal data and OCEL. Fig. 8a shows the ViEnNa model with structured visual storytelling, Fig. 8b presents the conventional oc-BPMN notation as a symbolic process diagram, and Fig. 8c depicts a free-form comic-like representation created by AI experts without structural modeling support.

The creation of ViEnNa model from Fig. 8a involved converting the OCEL 1.0 logs provided by Solve4X into the OCEL 2.0 format, followed by the processing using the ViEnNa framework. ViEnNa model from Fig. 8a goes as follows. The process begins with a state of no action until a client enters. Upon the client's entry, interactions occur simultaneously, and if there is no client, the process returns to the no-action state, repeating this cycle. Once a client is present, the next phase involves sequentially picking an asset and conducting a search, repeating this loop until the desired asset is found. If the asset is not found, the process reverts to the search phase. Upon finding the asset, both the client and the evaluator assess its quality. If the client is satisfied with the asset, they leave; if not, the process returns to searching. After the client departs, the system checks for a new client and repeats the entire procedure until the evaluation cycle concludes.



(a) ViEnNa model created over test multimodal data and OCEL.



(b) Conventional process model (oc-BPMN) created over the same test multimodal data and OCEL.

(c) Non-conventional (comic-like) models examples created by AI experts without using the ViEnNa framework.

Figure 8: Comparison of process modeling approaches over the same multimodal data and OCEL from Fehrer et al. (2024): (a) ViEnNa model, (b) conventional oc-BPMN model, and (c) non-conventional comic-like model created by AI experts.

E Evaluating the Prototype Implementation of ViEnNa

Scalability. We demonstrated that ViEnNa scales linearly with process complexity: the number of activities and evidence items increases processing load predictably. AI context-window limitations may emerge in large models, but style continuity can be maintained by teaching with a representative sample of prior images rather than full sequences.[2]

Contingency. ViEnNa operates gracefully in offline or low-resource settings. If GPU support is absent, panels render as static images and semantic search falls back to a cached approximate nearest-neighbor (ANN) index.[3] This results in the loss of live in-canvas highlights and incurs a 2–5$\times$ latency penalty, but enables uninterrupted offline model browsing and annotation.

Performance. For generation and retrieval tasks, ViEnNa responds in 3–7 seconds per image and under 10 seconds for a short (5 s) video snippet in ViEnNa-Motion.[4] These operations are processed server-side, providing consistent user experience across client platforms.

Cost and Deployment. Current deployments rely on server-hosted APIs for inference, but the full pipeline can be containerized and executed entirely offline. This flexibility supports integration in regulated or air-gapped environments, where no data leaves the local infrastructure.

Privacy and Fairness. All ViEnNa modules can operate locally without uploading sensitive data. Redaction layers automatically blur faces, mask personally identifiable information (PII), and log evidence retrievals. Fairness tests using gender-neutral variants showed no systematic bias in ICE-T scores;[5] ViEnNa adheres to OCEL object permissions and supports AES-GCM encryption for compliance with GDPR and HIPAA.

[2] In multimodal generation, context length constraints in large language models can lead to partial loss of narrative coherence unless representative priors are explicitly taught.

[3] ANN methods enable efficient similarity search in embedding space without recomputing pairwise distances.

[4] Processing time includes encoding, object referencing, and rendering depending on complexity and concurrency.

[5] Bias mitigation in generative models remains an open challenge, especially when training data lacks cultural and demographic diversity.

Summary. ViEnNa is deployment-ready for small-to-medium scale scenarios and privacy-conscious use cases. While not intended to be a standalone replacement for BPM platforms, it offers strong value as a narrative extension that emphasizes engagement, explainability, and multimodality.