MUMOSA, Interactive Dashboard for MUlti-MOdal Situation Awareness

Stephanie M. Lukin[1], Shawn Bowser[1], Reece Suchocki[1], Douglas Summers-Stay[1], Jeffrey Micher[1], Cynthia Matuszek[2], Francis Ferraro[2], Clare R. Voss[1]
[1]DEVCOM Army Research Laboratory, Adelphi MD, USA
[2]University of Maryland, Baltimore County, MD, USA
Correspondence: stephanie.m.lukin.civ@army.mil

Abstract

Information extraction has led the way for event detection from text for many years. Recent advances in neural models, such as Large Language Models (LLMs) and Vision-Language Models (VLMs), have enabled the integration of multiple modalities, providing richer sources of information about events. Concurrently, the development of schema graphs and 3D reconstruction methods has enhanced our ability to visualize and annotate complex events. Building on these innovations, we introduce the MUMOSA (MUlti-MOdal Situation Awareness) interactive dashboard that brings these diverse resources together. MUMOSA aims to provide a comprehensive platform for *event situational awareness*, offering users a powerful tool for understanding and analyzing complex scenarios across modalities.

1 Introduction

After a significant incident or crisis, how do *investigators* determine and assess what happened when in order to produce a report with clear evidence from the sequence of events in detailing lessons learned? How do communities prepare their *responders* to handle similarly complex, critical situations that may come their way in the future? Some crisis response procedures are well-established for specific situations, e.g., an initial fire suppression response to a wildfire[1], so responders can be consistently trained in advance and investigators know what to look for afterwards. But other times, the crisis is so sudden and unexpected that established lines of communication struggle to convey up-to-date information. Following these unforeseen circumstances, both groups, investigators and responders, have a shared need to understand the various types of information about events in the evidence collected and analyzed for post-crisis reports.

The field of information extraction (IE) within computational linguistics has led the way since the late 1980's applying symbolic, then statistical, and most recently neural methods to natural language texts to identify the types of essential elements of information needed for such reports, including entities, relations, and events (Grishman, 2019). Most recently, with neural models such as LLMs and VLMs that can bring together multiple modalities to provide additional sources of information about events, there is now the opportunity to leverage various combinations of multimodal event information to support investigators in combing through text and photographic evidence for report writing and to train responders in preparing to handle such information in the future. Furthermore, the recent development of schema graphs with access to over 3K event types (Zhan et al., 2023) and 3D reconstruction methods for scenario simulation from as little as 24 images, e.g., Kerbl et al. (2023), users can now have hands-on access to interfaces to visualize and annotate complex events online, as they learn from available evidence and documentation what has happened over the course of those events.

In this paper, we introduce our approach to bringing together these various resources in an interactive, MUlti-MOdal Situation Awareness (MUMOSA) dashboard for complex event understanding, ultimately in support of users' real-time event *situational awareness* (SA) and decision-making during a crisis. For a specific role, such as a first responder onsite or incident coordinator at an emergency operations center, the specifics of their SA will be determined by tasks and decisions for their job. However in all cases, their SA will entail "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future." (Endsley, 1995, 2015)[2]

[1]https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fsm9_039213.pdf

[2]Perception, comprehension, and projection have been

Proceedings of the Workshop on the Future of Event Detection (FuturED), pages 32–47

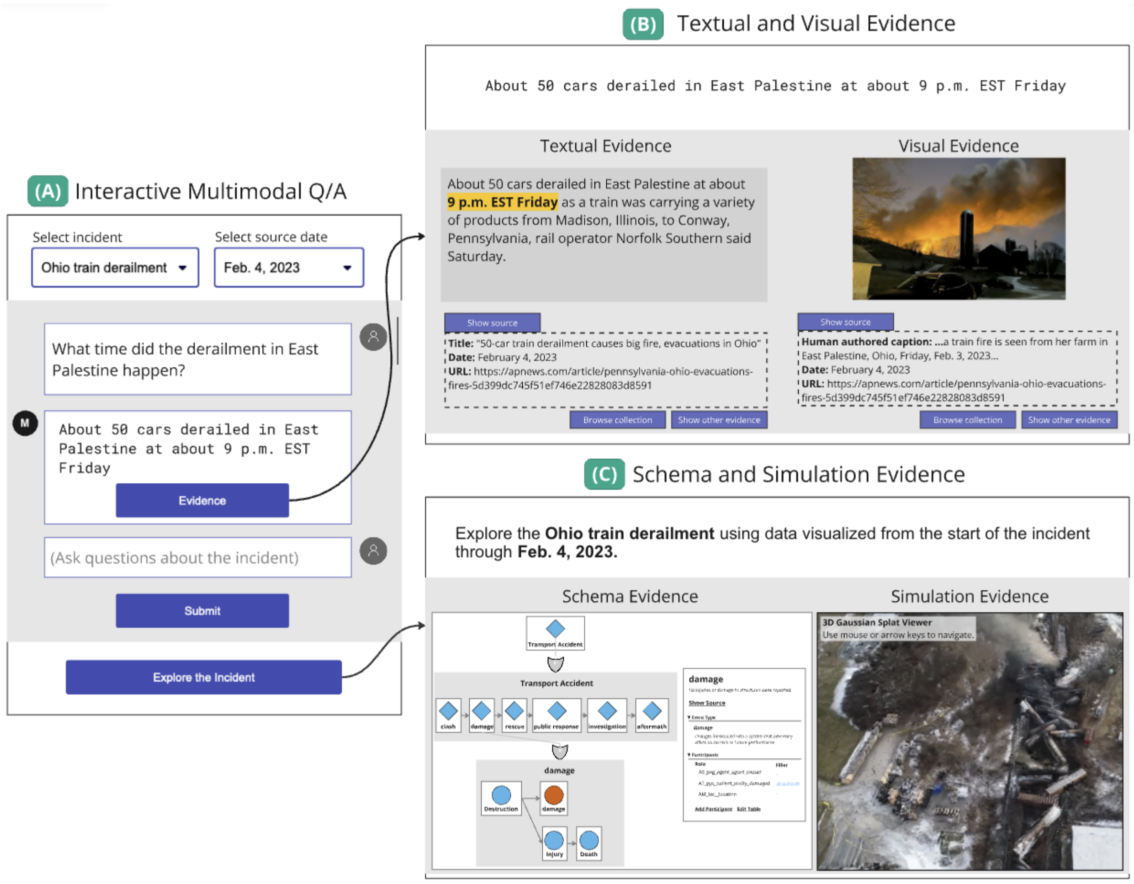

Figure 1: Users engage with the MUMOSA dashboard through (**A**) Interactive Multi-Modal Q/A, selecting the incident with source date and entering questions or directed requests in natural language. They may then choose to examine the source data for the system response via panels in (**B**) Textual and Visual Evidence, or continue engaging the dashboard via Q/A. The selected incident can also be explored in its entirety across events via panels in (**C**) Schemas and Simulation Evidence.

The MUMOSA dashboard's panels are intended to provide for user perception of essential elements of information (SA level 1) about the complex event they select in the form of multi-modal evidence. The workflow design enables users to explore and compare information across evidence panels, as well as edit and annotate the content of the complex schema graphs and scenario simulation. This aims to support them in interpreting and retaining the panel content with multiple elements of information, and in building their own narrative of the complex event (level 2).

Users engage in interactive multi-modal question-answering (Q/A) and exploration of events and simulated environments. The user initializes the dashboard for the complex event of interest and time frame. After inputting a question, they receive a text answer and access to supporting evidence from text documents and photographic visuals. They may simultaneously explore the event in its entirely through schema graph structures and 3D simulations. Each of these modalities of evidence pertaining to the incident appears in a separate interactive panel, as shown in Figure 1.

In this paper, we select one unexpected crisis to focus on—the Ohio 2023 train derailment—as we describe the design and capabilities of the dashboard for the following applications:

1. to assemble *crisis documentation* for those writing 'lessons learned' investigative reports

2. to create *training resources* for those responsible in the future for handling crises

For background, Section 2 describes existing resources we build on in constructing the dashboard, and basic facts about the Ohio derailment. Sections 3 and 4 cover the dashboard implementation and workflow (SA levels 1 and 2 respectively) for the crisis documentation and training resources. Section 5 envisions the dashboard of the future sup-

designated as progressive levels 1, 2, and 3 of SA.

porting real-time crisis response, akin to the needs of SA level 3. We briefly overview related research that differs from our approach in Section 6 and conclude in Section 7.[3]

2 Background

We briefly overview here existing resources that we build on in four panels of our dashboard.

Textual evidence. To show textual evidence, we leverage text-based Q/A and Frequently Asked Questions (FAQs), where a set of common user questions pertaining to a particular topic are compiled into an accessible list where the user may look up answers if their query is common (Tekumalla, 2020). Prior work has shown how to find these matches in a dialogue Q/A. By using statistical text classifiers, Leuski and Traum (2011) compared a user question in real-time against a distributions of common user questions paired with responses. After successfully matching the input question to the pre-processed question set, the paired answer was returned. The same statistical classifier was leveraged in human-robot dialogue extending beyond the constraint of a 'question,' allowing for different types of frequently issued robot-directed commands (Lukin et al., 2018; Gervits et al., 2021). The matching was re-implemented in Lukin et al. (2024) using sentence embeddings and cosine similarity to find close matches between vectorized input and pre-stored questions, and showed significant improvement in accuracy over the statistical approach in the same domain of human-robot dialogue. Section 3.1.1 covers our work incorporating a modified version of this approach with LLMs in responding to user input.

Visual evidence. Prior to VLMs, comprehensive text generation from images required several different tools: OCR for text recognition, object recognition and segmentation for object annotation, visual-question-answering models for short answers to specific questions about the image, and captioners to generate a one sentence description of the image. Now, VLMs are able to accomplish all of these tasks (with the present exception of producing segmentation and bounding boxes) in a unified, context-sensitive way. To show visual evidence in answering a text question, Section 3.1.2 describes our use of Idefics3 (Laurençon et al., 2024) to generate text descriptions of images that we can match on in a similar way as carried out for textual evidence. Idefics3 was developed by Huggingface and builds off of Google's SigLIP (Zhai et al., 2023) and Meta's Llama 3.1 (AI@Meta, 2024).

Schema evidence. Schemas provide structured representations of real-world occurrences. They are event-centric, and as such, serve as abstract templates for understanding and analyzing complex sets of events. Event schemas typically consist of:

- Events: High-level (e.g., "transport accident") and granular sub-events (e.g., "damage," "investigation.")
- Entities: Actors or objects involved in events (e.g., "train," "residents," "authorities.")
- Relations: Connections between entities or events, often temporal or causal in nature.

Schema visualizers allow for complex events to be viewed in an intuitive way via a graph-like structure of nodes with directed edges. Schemas may be compared against source documents to find when an event mentioned matches an event node. Section 3.2.1 describes the RESIN pipeline are used to extract and match events to the schema (Du et al., 2022; Wen et al., 2021) and the RESIN visualizer (Nguyen et al., 2023) to view and edit them within our dashboard.

Simulation evidence. Reconstructing scenes from a set of images is an emerging research area in 3D computer vision, enabling novel view synthesis and embodied scene understanding, both of which could be crucial for crisis response. Advancements in Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) have resulted in 3D models that achieve state-of-the-art in rendering appearance, rendering speed, and training efficiency. Section 3.2.2 covers the use of 3DGS (Kerbl et al., 2023). A 3DGS scene is represented as a set G of discrete Gaussian primitives each with parameters (μ, o, s, r, SH) where $\mu \in \mathbb{R}^3$ is for the spatial center, $o \in \mathbb{R}$ for opacity, $s \in \mathbb{R}^3$ for scale, $r \in \mathbb{R}^4$ for quaternion rotation, and SH for spherical harmonics coefficients which represent view-dependent colors. We extend this formulation with additional parameters to capture language features from 2D foundation models (Kirillov et al., 2023; Bowser and Lukin, 2024).

Scenario: Ohio Train Derailment. On February 3, 2023, in East Palestine, Ohio, USA, about 50 train cars derailed from a 150 Norfolk Southern

[3]We plan to release the dashboard as an artifact of our research towards enabling higher levels of SA following systematic evaluation in future work.

freight train.[4] Eleven of the derailed cars were carrying hazardous materials including vinyl chloride, ethylene glycol, ethylhexyl acrylate, butyl acrylate and isobutylene. Some cars caught fire, and others spilled hundred thousand gallons of hazardous materials into a stream that eventually empties into the Ohio River. A number of federal and state government agencies were immediately mobilized. Clean-up efforts included real-time testing of air, soil, and water. A controlled burn of remaining chemicals was ordered on February 6, 2023 to prevent further explosions. However, after extensive investigation of the incident, this course of action was assessed over a year later to have been unnecessary. We select this real-world incident since news articles, government reports, photos, and other data about it are openly available for populating our dashboard panels and assessing ways these varied information sources may help immediate responders more reliably gain SA in novel and unexpected events.

3 MUMOSA Dashboard Evidence Panels

The user starts their interactions with the MUMOSA dashboard by initializing it for a particular incident and date of interest. They can then proceed by posing questions about events in the incident and exploring the source data evidence provided with the system responses along with full incident visualizations in the dashboard panels. We focus here and in Section 4 on interactions specifically for *forensic* use cases of the dashboard with pre-processed data for *investigators* and *first responders.* We postpone till Section 5 discussion of future dashboard research for *real-time* conditions with dynamic changes to both data availability and user information needs.

The intent for the dashboard panels is to provide users with essential elements of information (SA level 1). User workflow across panels for building their understanding of the sequences of events within the incident (SA level 2) will be addressed in the section that follows.

3.1 Interactive Multi-Modal Q/A

During the multi-modal Q/A interaction, the user enters questions or makes directed requests using natural language, and the dashboard responds, as information is available, with both a text answer (Figure 1**A**) and panels populated with supportive evidence from source texts and visuals (Figure 1**B**). The system provides next-search alternatives to Q/A interactions for the users within the panels, enabling them to look for other evidence deeper within the reply stack or browse the source document collection.

3.1.1 Textual Evidence Panel

Textual evidence is shown in its own panel with the system answer (the text returned to the user in response to their question in the Q/A interaction) highlighted, and surrounding context and source information provided for further exploration. *In advance of the user's question*, the dashboard contains a collection of texts. For our case study scenario, we gathered news articles from different sources published on different days following the derailment. The text from these articles was scraped from the websites and segmented into sentences. Next, we created Q/A pairs, where the questions had answers contained in the sentences. We ran different large language models (ChatGPT and Llama 3.1) to generate numerous questions from the sentences, and then with manual review, as feasible for this forensic use case, we validated or adjusted each generated question as reasonable for inclusion in our stored Q/A pairs. For subsequent run-time comparison with user questions, all the stored questions were also vectorized through sentence embeddings using SBERT (Reimers, 2019).

After the input of the user's question, we run a semantic sentence matching, as described in Lukin et al. (2024). The user question is vectorized with SBERT and compared against every pre-stored vectorized question using cosine similarity. The stored questions are ranked in descending order, and the answer, found in the pair with the top question, is displayed to the user in the answer box immediately below the user question. The answer also appears contextualized within the Textual Evidence Panel containing: the source document passage with answer highlighted, the document title, link to the source document, and two buttons for further exploration. The "Show other evidence" button enables users to examine other lower-ranked answers retrieved in response to their request. The "Browse collection" button gives users access to the system's text document collection for more extensive investigation of source materials. Table 1 shows the user question *"What time did the derailment in East Palestine happen?"* as matched against the top-3 answers, corresponding to a request to show other evidence in Figure 1**B**.

[4]EPA website on derailment: https://www.epa.gov/east-palestine-oh-train-derailment/operational-updates

Table 1: Ranked answers to user question: *"What time did the derailment in East Palestine happen?"*

Score	Answer
0.82	About 50 cars derailed in East Palestine at about 9 p.m. EST Friday...
0.71	The 50-car Norfolk Southern train derailed around 9 p.m. Friday night.
0.53	East Palestine officials said 68 agencies from three states and a number of counties responded to the derailment...

3.1.2 Visual Evidence Panel

Visual evidence is shown in its own dashboard panel in response to the user's question. In *advance of the user's question*, the dashboard requires a collection of images paired with natural language text. This text may come from different sources, including:

- human-written captions or alt-texts associated with the image, if retrieved from a document;
- machine-generated captions or descriptions as generated by a VLM;
- annotated labels or bounding boxes associated with objects in an image as annotated by a computer vision object detection model;
- texts within an image generated by OCR.

For our forensic use case in the case study, we manually gathered photographs from different sources showing the derailment. We then extracted the relevant text sources (i.e., the caption, alt-text, OCR) and generated text descriptions using Idefics3-8B-Llama3.

After the input of the user's question, we follow a paired-vector comparison process, similar to the one described for textual evidence: the user's question is vectorized using SBERT and compared against pre-stored, vectorized image texts. The pair with the highest match is selected, the matched image text is returned to the user as the answer to their question, and the associated image is shown in the Visual Evidence Panel alongside the image texts. For example, a different user question inquiring into the aftermath of the derailment might be, *"How does the wastewater get cleaned?"* This question might have a high match to the human-authored text caption "This centrifuge separates solid waste from liquid wastewater in holding bins to determine whether the waste is hazardous and then disposed properly" that is associated with an image of a centrifuge. This caption is returned to the user as the answer, and the retrieved image will appear together with its text caption in the Visual Evidence Panel (Figure 2). The user may follow up with two button choices, asking the system to show other visual evidence found, or browsing the source collection of images to inspire new questions.

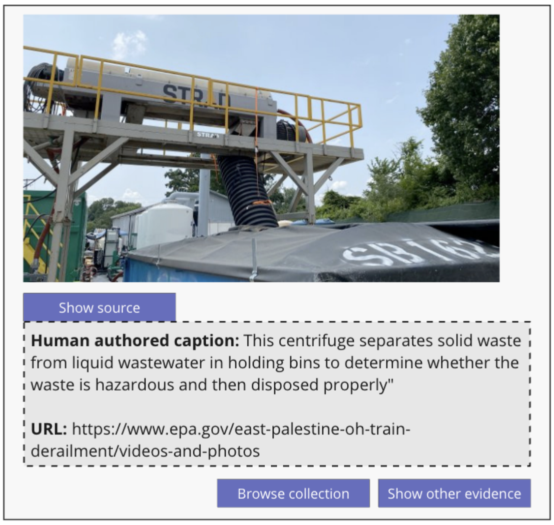

Figure 2: Visual Evidence Panel following the user question *"How does the wastewater get cleaned?"*

In cases where the user's question yields a high match from both the collection of documents and images, the dashboard will inform the user through the Q/A interface to inspect both Textual and Visual Evidence Panels.

3.2 Interactive Multi-Modal Event Exploration

In addition to asking questions or making directed requests, the user can explore the event as a whole using two representations of the incident: a schema panel view showing a graph of event nodes and either hierarchical or temporal event-event relations as edges, and a 3D simulated panel view showing a visualization of event itself. The Schema and Simulation Evidence Panels are not extrinsically tied to the user's question. At any point during the user's interaction with the dashboard, they may choose to explore these panels.

3.2.1 Schema Evidence Panel

The Schema Evidence Panel provides the user with an event-centric exploration of the incident. First, documents are preprocessed to identify events using the information extraction module of the RESIN pipeline (Du et al., 2022; Wen et al., 2021). The resulting extractions are then matched to events in incident schemas by executing the matching module in the RESIN pipeline. Finally, a separate program consumes the matching module's output for visualizing the schema graph in the evidence panel where the user can explore and edit it (Nguyen et al., 2023).

The schema the user sees is based on their selection of the incident and timeframe at the top of the dashboard. Figure 1**C** shows a screenshot of the *transport accident* schema[5], and so the visualized graph displays events within this type of incident. Blue diamonds represent complex events in the schema that typically happen or could happen within this incident. Clicking a blue diamond, e.g., *damage*, expands the graph with its subevents containing different color and shaped nodes. Circles are primitive events, as leaf nodes (no subevents). A red circle indicates actual evidence was extracted from the source material that matched the primitive event type, whereas events not included in the original schema appear as yellow (see Figure 4 in Appendix A).

Clicking a red circle, e.g., *damage*, expands the panel with further information about the event, including the matched phrase in the source material, and the participants and their roles in the event (i.e., A0 agent causer, A1 patient entity damaged, and location). The user has hands-on access to explore all events within the schema, opening nodes to see where matches occurred in event fields from document content reported during the selected timeframe.

3.2.2 Simulation Evidence Panel

The Simulation Evidence Panel provides the user with an interactive 3D model that has been reconstructed from photographs of the incident, resulting in a bird's-eye view of the scene. The user can navigate the scene through the first-person perspective using a mouse and keyboard.

The simulation is constructed using a 3D Gaussian Splatting point cloud structure (Kerbl et al., 2023). Figure 1**C** shows a snapshot from our simulation after scene reconstruction from the angle the user selected by moving their cursor. This simulation was created using only 24 references images that were captured by an aerial drone flying overhead of the derailment incident,[6] thus the reconstruction shows the user novel views unavailable from the original source. Figure 7 in Appendix B shows the full flattened view of the 3D simulation from which the view in Figure 1**C** was taken. The geometry and visual appearance of the simulation is improved as more images are added.

The simulation can be annotated using user-specified keypoints and image segmentation masks which are unprojected onto the underlying model for 3D segmentation. Sections of the 3D map can then be highlighted with unique colors and icons.

4 User Workflow

The MUMOSA dashboard represents a powerful tool for users to query events by enabling Q/A over multi-modal data sources where the modalities offer complementary supporting evidence. While the evidence in the individual panels provides users with essential elements of information (SA level 1), the MUMOSA dashboard itself provides for an easy-access workflow to detect and compare events across panels and modalities. By using all the panels together, the dashboard provides unique opportunities for users to iterate in their information foraging and annotate the underlying data to enhance their understanding of the sequence of events (SA level 2).

In our forensic use of this case study, a user may want to understand the initial response of the derailment by asking the question, *"What time did the derailment in East Palestine happen?"* to which the answer was *"about 9 p.m. EST Friday"* with the textual evidence showing an article published on February 4, 2023. This answer may prompt the user towards several lines of inquiry, one of which may be to ask *"Was it hard to see at night during the initial incident response?"* This might return a photograph of the nighttime scene in the Visual Evidence Panel, which may in turn inspire another question that could be answered by the Textual Evidence Panel, e.g., *"What challenges did the first responders face in the dark?"*

We also readily imagine that the open-ended workflow with schema and simulation panels will elicit follow-up questions. When exploring the schema view, users can navigate the hierarchical structure of events related to the train derailment. For example, they might explore sub-events under *investigation*, as seen Figure 6 in Appendix A. This exploration may prompt a user to follow-up with *"What criminal charges are being reported?"* The response in the schema evidence would highlight the node of interest within the schema, and search

[5]The *transport accident* schema was independently curated by RESIN team on DARPA KAIROS.

[6]Source video from Youtube: "National Transportation Safety Board B-Roll: Train Derailment in East Palestine, OH" https://www.youtube.com/watch?v=7AyXTVkVBT4

stored documents for supportive textual evidence.

Giving users access to the simulation affords them greater situational awareness to ask questions pertaining to accessibility and route planning, obstacle and target identification, and scene overview. Examination of the simulation might prompt not only new questions, but the ability to display answers within the simulation using annotations. The user may ask the question *"Where is the immediate danger?"* and the simulation would highlight the clusters of train cars in orange and red, as seen in Figure 3. This may be followed up by *"What buildings are in immediate danger?"* and would highlight the buildings in green (also Figure 3). In this way, the simulation view displays portions of visual evidence to the user without constraining their viewpoint to the original camera pose.

Figure 3: Simulation evidence augmented with semantic segmentation masks as a result of user questions

5 Discussion: Toward Real-Time Event Tracking

The MUMOSA dashboard currently aims to serve as an interactive *forensic resource,* providing support to post-crisis incident investigations and training exercises for first responders at SA levels 1 and 2. The questions we have included above showcase how users may search for information looking back at events within an incident across modalities. The evidence supplied in one modality may inspire new questions or may lead to further insights in conjunction with evidence from another modality. With this groundwork in place, we now shift our discussion to how we envision the dashboard will support *real-time* crisis responses.

5.1 Dynamic Timeframes

The MUMOSA dashboard is designed to show grounded evidence for the incident and timeframe the user selects at the start of a session to build their own understanding of time-stamped incident events. It remains an open design research question how we might modify the dashboard to *automatically visualize* incident changes for the user in real-time, without also cognitively overloading them by viewing too much information across the multiple modalities. One UI/UX design opportunity is augmenting the dashboard with a timeline and adjustable slider for the user to control the sequenced, connected display of photos, news report summaries, 3D reconstructions and schema graphs. Photographs from news reports of ongoing events could be presented along a timeline to show the progression of events together with generated text summaries based on and time-aligned to those reports for augmenting a Situation Report, such as SmartBook (Reddy et al., 2024). The 3DGS simulation within the Simulation Evidence panel, in conjunction with an adjustable slider on the timeline, could display changes to the simulation by adjusting the opacity of Gaussians belonging to dynamic objects (Shen et al., 2024; Wu et al., 2024). Similarly, the slider could be connected to the graph display in the Schema Evidence panel, enabling the user to move through the progression of photos on the timeline in conjunction with visible changes to the schema, displaying automated detection of events in red graph nodes (Appendix A shows different timescales of schema evidence).

Additional modalities may extend to time series data collected at the incident site from sensors deployed that are constantly recording and storing measurements. In particular, we are exploring how time series data from the air[7], water[8], and soil sample measurements[9] can be incorporated into the dashboard to allow a user to examine quantitative data changes over time and location and further query these new modalities through the Q/A interaction.

5.2 Scalability for Responding to Rapidly Evolving Incidents

In order for the dashboard to be responsive to real-time event tracking, the back-end storage and processing requirements must be scalable to support streaming data, as well as filter the incoming data for content, such as for misinformation. Though these issues fall beyond the scope of this paper, they

[7]https://www.epa.gov/east-palestine-oh-train-derailment/air-sampling-data

[8]https://www.orsanco.org/east-palestine-train-derailment-spill-response/

[9]https://www.epa.gov/east-palestine-oh-train-derailment/soil-and-sediment-sampling-data

help sharpen the criteria and distinctions to keep in mind as we are in the process of selecting metrics and designing an evaluation of the current MUMOSA dashboard with its strictly forensic goals. For example, the intended end-users of the forensic MUMOSA dashboard will not be subject to the time pressure, cognitive distractions, and levels of noise in an emergency operations center or an incident command post that end-users of a real-time dashboard would be. The speed of processing (velocity), the amount of data (volume), the range of data modalities (variety), the timeliness and accuracy of the data content (value and veracity)—all well-known "V"s of information overload—will differ along with stakeholder and user expectations and requirements of a dashboard, depending on whether it will serve forensic or real-time goals. These all need to be understood in advance of any dashboard evaluation.

5.3 Evaluating User Priorities

Ultimately, our goal is to support end-users of a real-time dashboard at SA levels 1, 2, and 3, that includes scenario planning and "what-if" analysis using all available modalities. In determining how effective the dashboard can be for investigators and responders at SA levels 1 and 2, we have begun to assess the accuracy of the technology in each evidence panel. Before we can measure how well the dashboard can support different user needs, as we are not subject matter experts for their tasks, we need to design and conduct interviews with individuals in the relevant communities.

Investigators, who examine past events with particular questions in mind, will benefit from training on the dashboard before any evaluation, with guided learning of panel workflows and in-depth searches that support chronological reconstruction (such as browsing the document collection and schema visualization over time). This already suggests additional value to prioritizing the development of the timeline mentioned above for a real-time dashboard. Furthermore, a future dashboard that provides automated detection and highlighting of discrepancies between conflicting event reports would also help expedite the investigator's work. For now, we plan to task participants involved in our *investigator evaluation* with manually constructing a timeline of events and their trusted sources, to assess the ease with which they can make use of the current dashboard.

Responders, who are in training to deal with crisis events, will benefit from learning to view the 3D simulations and annotations that document immediate dangers in physical environments, for discerning what constitutes sufficient information for rapid decision-making. Thus, a future dashboard that provides for automatic detection and highlighting of key crisis regions in the physical environment as the incident unfolds, would help expedite the responder processing of incoming information. For now, we are designing a pilot study for participants in the *responder evaluation* where their task will be to create a situation report with summaries of levels of danger at different locations, as more information becomes known over time.

5.4 Event Tracking with a Dialogue Agent

The dashboard as presented thus far relies on the user to "pull" information by typing and submitting their own questions as inspired by their viewing of the different evidence panels. We envision a future iteration where the dashboard becomes more like an *agent* that can also "push" information to the end-user, taking an active role in the interaction. The agent could engage in a dialogue with the user by supporting a fact-seeking conversation, requesting clarifications to the user's question, and making suggestions of follow-up questions. The agent may be designed to recognize changes to a situation as time progresses, and alert the user so they may further inquire.

Additionally, we expect to extend the dashboard beyond text input only to support various modality input, where the user could submit text questions about particular image(s) previously retrieved as visual evidence. We envision an additional interaction mode within the dashboard that can reference prior images and engage in a new dialogue session about them. This further would support new images to be uploaded as incidents unfold in real-time. In these cases of unlabeled data, human-authored captions may not be available, and instead, automatically generated descriptions can be used for Q/A matching. Table 2 shows both the human-authored descriptions about the image in Figure 2, as well as automatically generated ones, which provide extensive detail and attempted analysis.

6 Related Work

Text-based search methods, begun in the 1950's, have been central to modern information retrieval systems (Sanderson and Croft, 2012). Most recently, generative AI has been leading a new

Human-written caption	"This centrifuge separates solid waste from liquid wastewater in holding bins to determine whether the waste is hazardous and then disposed properly."
Alt-text	"Huge machine called a centrifuge separates the solid waste."
Text in image	"STRAD" and "SB162"
Generated description from Idefics3-8B-Llama3 (Prompt: "Describe this image")	"The object is most likely a type of pump station designed for heavy-duty use, probably utilized in industries like construction, mining, or wastewater management due to its robust build and extensive piping setup. Pumps labeled with identifiers like SB168 often denote standardized models within these sectors, known for their reliability and durability. Given the context clues provided by the surrounding environment (industrial equipment, trees), we infer that this equipment plays a critical role in managing fluid flow efficiently across different sites where liquid transfer is necessary." (See Appendix C for the full generated output)

Table 2: Texts associated with the photograph retrieved in Figure 2

wave of search-related technologies (Najork, 2023). Nonetheless, it is quite striking that reports suggest half of all web searches are not answered.[10] Thus, for complex tasks such as those build situational awareness, there remain significant opportunities to develop, deploy, and assess interactive systems for the impact of providing multiple modalities of information to dashboard users for their detection and understanding of events over time.

We know of no other research that combines information for user access across modalities as we have within an interactive dashboard. MUMOSA provides an interactive, multi-modal interface where users can iteratively forage for answers about complex events to meet their information requirements. Our approach, by retaining a dialogue history with text and visual evidence for documenting user searches, paves the way to building an AI agent-based system (White, 2024).

Event detection has recently been expanded by novel methods of embedding and extracting events across modalities from multimedia sources. For example, by constructing shared semantic vector spaces for texts and images (Radford et al., 2021; Jia et al., 2021), systems can generate text descriptions of events detected in images where only objects have been identified (Li et al., 2020a). Image retrieval has seen advances by using global features (Shao et al., 2023), augmenting query or image vectors (Zhu et al., 2023), and general purpose VLMs (Wang et al., 2022). Our FAQ approach on documents and images is intended to achieve high accuracy to support our investigator users looking forensically at data, and serve as a strong feasibility test in bringing together the evidence panels. We also give users access to the document and image collections using the semantic search ranking to enable users to forage in a less constrained way by exposing the evidence directly to the user.

The automated construction of regular patterns of events from news reports, where the task of event schema induction applies, continues to challenge researchers (Devare et al., 2023; Li et al., 2023). The innovation of building path language models by connecting shared arguments across events within instance graphs has provided for more complete schema induction (Li et al., 2020b).

Recent advances in 3DGS for 3D reconstruction have begun to support natural language querying of a scene (Shi et al., 2024), however these approaches presently only highlight segmentation masks based on keywords, e.g., 'train cars.' There is no framework in place to support interactive querying of the 3D reconstruction from natural language questions. In order to understand that the question we pose *"Where is the immediate danger?"* refers to the train cars requires greater understanding of 'danger' in the context of the simulation.

7 Conclusion

Our MUMOSA dashboard aims to provide a user with level 1 and 2 situational awareness for understanding essential elements of information and complex events by uniting complimentary modalities and interactions. We further envision how the dashboard will support real-time crisis response (SA level 3). By integrating document-based Q/A, visual evidence retrieval, event schema visualization, and 3D scene simulation, our dashboard offers a comprehensive solution for complex event understanding. This multifaceted approach not only supports various levels of situational awareness, from initial perception to comprehensive understanding, but also provides a flexible, future-ready framework that can evolve with advancements in AI and data processing technologies.

[10]https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web

References

AI@Meta. 2024. Llama 3 model card.

Shawn Bowser and Stephanie M. Lukin. 2024. 3D Gaussian Splatting for Human-Robot Interaction. In *Interactive AI for Human-Centered Robotics Workshop*.

Sugam Devare, Mahnaz Koupaee, Gautham Gunapati, Sayontan Ghosh, Sai Vallurupalli, Yash Kumar Lal, Francis Ferraro, Nathanael Chambers, Greg Durrett, Raymond Mooney, et al. 2023. Sageviz: Schema generation and visualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 328–335.

Xinya Du, Zixuan Zhang, Sha Li, Pengfei Yu, Hongwei Wang, Tuan Lai, Xudong Lin, Ziqi Wang, Iris Liu, Ben Zhou, Haoyang Wen, Manling Li, Darryl Hannan, Jie Lei, Hyounghun Kim, Rotem Dror, Haoyu Wang, Michael Regan, Qi Zeng, Qing Lyu, Charles Yu, Carl Edwards, Xiaomeng Jin, Yizhu Jiao, Ghazaleh Kazeminejad, Zhenhailong Wang, Chris Callison-Burch, Mohit Bansal, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, Martha Palmer, and Heng Ji. 2022. RESIN-11: Schema-guided event prediction for 11 newsworthy scenarios. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 54–63, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Mica R. Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Journal of the Human Factors and Ergonomics Society*, 37(1):32–64.

Mica R. Endsley. 2015. Situation awareness misconceptions and misunderstandings. *Journal of Cognitive Engineering and Decision Making*, 9(1):4–32.

Felix Gervits, Anton Leuski, Claire Bonial, Carla Gordon, and David Traum. 2021. A classification-based approach to automating human-robot dialogue. Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representa- tion learning with noisy text supervision. In *Proceedings of the 38 th International Conference on Machine Learning, PMLR 139*.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *Preprint*, arXiv:2304.02643.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *Preprint*, arXiv:2408.12637.

Anton Leuski and David Traum. 2011. Npceditor: Creating virtual human dialogue using information retrieval techniques. *Ai Magazine*, 32(2):42–56.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020a. Cross-media structured common space for multimedia event extraction. In *Proceedings of ACL*.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020b. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023. Open-domain hierarchical event schema induction by incremental prompting and verification. *Preprint*, arXiv:2307.01972.

Stephanie M Lukin, Felix Gervits, Cory J Hayes, Anton Leuski, Pooja Moolchandani, John G Rogers III, Carlos Sanchez Amaro, Matthew Marge, Clare R Voss, and David Traum. 2018. Scoutbot: a dialogue system for collaborative navigation.

Stephanie M. Lukin, Jaci South, and Shawn Bowser. 2024. CHRIS-Bot: A Robot for Dialogue and Scene Understanding of Anomalous Environments in Virtual Reality. Technical Report ARL-TR-9906, DEVCOM Army Research Laboratory.

N Najork. 2023. Generative information retrieval. In *Proceedings of the 46th Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval*, volume 1.

Khanh Duy Nguyen, Zixuan Zhang, Reece Suchocki, Sha Li, Martha Palmer, Susan Brown, Jiawei Han, and Heng Ji. 2023. Resin-editor: A schema-guided hierarchical event graph visualizer and editor. *Preprint*, arXiv:2312.03093.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38 th International Conference on Machine Learning, PMLR 139*.

Revanth Gangi Reddy, Daniel Lee, Yi R. Fung, Khanh Duy Nguyen, Qi Zeng, Manling Li, Ziqi Wang, Clare Voss, and Heng Ji. 2024. Smartbook: Ai-assisted situation report generation for intelligence analysts. *Preprint*, arXiv:2303.14337.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Mark Sanderson and W. Bruce Croft. 2012. The history of information retrieval research. In *Proceedings of the IEEE, Issue: Special Centennial Issue*, volume 100, pages 1444–1451.

Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. 2023. Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046.

Licheng Shen, Ho Ngai Chow, Lingyun Wang, Tong Zhang, Mengqiu Wang, and Yuxing Han. 2024. Gaussian time machine: A real-time rendering methodology for time-variant appearances. *Preprint*, arXiv:2405.13694.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. 2024. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5333–5343.

Lavanya Sita Tekumalla. 2020. NLP Tutorial : Automatic Question Answering from information in FAQ. https://github.com/lavanyats/QuestionAnswering_From_FAQ_Tutorial. Published: 2020. Accessed: 2024-10-01.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Ryen W. White. 2024. Advancing the search frontier with ai agents. *Communications of the ACM*, 67(9):54–65.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. 4d gaussian splatting for real-time dynamic scene rendering. *Preprint*, arXiv:2310.08528.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Qiusi Zhan, Sha Li, Kathryn Conger, Martha Palmer, Heng Ji, and Jiawei Han. 2023. Glen: General-purpose event detection for thousands of types. *Preprint*, arXiv:2303.09093.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A Appendix: Schema Visualization and Sources

This section lists the four article sources used to create the different schema timeframes.

Article 1:

- Title: 50-car train derailment causes big fire, evacuations in Ohio
- Date Published: February 4, 2023
- Source: Associated Press (AP)
- https://apnews.com/article/pennsylvania-ohio-evacuations-fires-5d399dc745f51ef746e22828083d8591

Article 2:

- Title: East Palestine under mandatory evacuation, possible explosion warning after toxic train derailment
- Date Published: February 5, 2023
- Source: Ideastream Public Media
- https://www.ideastream.org/community/2023-02-05/east-palestine-under-mandatory-evacuation-possible-explosion-warning-after-toxic-train-derailment

Article 3:

- Title: Ohio crews conduct a 'controlled release' of toxic chemicals from derailed train cars
- Date Published: February 6, 2023
- Source: National Public Radio (NPR)
- https://www.npr.org/2023/02/06/1154760911/ohio-train-derailment

Article 4:

- Title: Is the Ohio River contaminated? East Palestine train derailment sparks concerns over water
- Date Published: February 14, 2023
- Source: The Columbus Dispatch
- https://www.dispatch.com/story/news/local/2023/02/14/is-the-ohio-river-contaminated-east-palestine-train-derailment-sparks-concerns-over-water/69900528007/

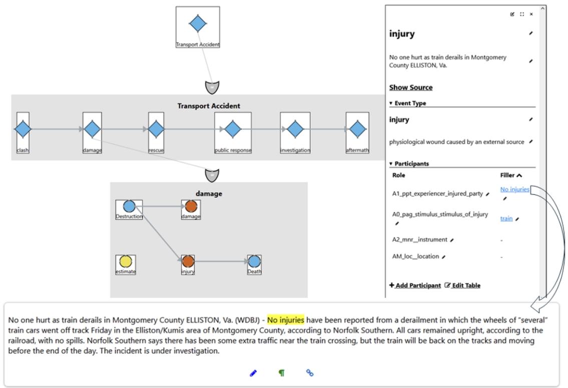

Figure 4: Schema Evidence Panel using the transport accident schema, matched on events in news reports over 10 days. Selecting 'injury' (red) node and matched text "No injuries" in role filler, opens panel below with that text highlighted in source context.

As more documents are analysed over time, additional extracted events are incorporated into MUMOSA. Some event nodes may be newly matched (previously blue, turn red), such as "injury", while newly extracted events not included in the original schema may appear (yellow), such as "estimate", as shown in Figure 4. Selecting the linked filler in the panel of a matched node opens another window containing the matched phrase in the context of the source material.

Figure 5 presents two schemas side by side, to make evident the contrast between earlier and later date information within same schema. Both schemas match and visualize on the "damage" event, first with extracted information from just the earliest document (Article 1), and then from all four articles over ten days. Figure 6 similarly presents two schemas corresponds to the schema matching and visualization on "investigation" events, first with extracted information from just the earliest document (Article 1), and then from all four articles over ten days.

B Appendix: Simulation Space

Figure 7 displays a flattened view of the 3D simulation modeled using 3D Gaussian Splatting from the source video: "National Transportation Safety Board B-Roll: Train Derailment in East Palestine, OH" [11]

[11] https://www.youtube.com/watch?v=7AyXTVkVBT4

C Appendix: Full Image Description Output

Table 2 abbreviated the generated output by defics3-8B-Llama3 with the input image in Figure 2 using the prompt "Describe this image". Below is the full output from the model, verbatim (** indicates bold formatting):

The image depicts a piece of industrial equipment situated outdoors, likely in an area designated for construction or maintenance work. This apparatus appears to be part of a larger system used for pumping liquids or slurries, given the presence of hoses and pipes connected to it

Description of Objects Present in the Image:

1. **Main Equipment**:
- A large machine with multiple components:
- It has a rectangular base that supports several vertical elements.
- There is a prominent horizontal pipe extending from one side of the main structure, which seems to have black corrugated tubing attached to its end. This suggests that the device might be involved in transferring fluids or materials through pipelines.

2. **Safety Features**:
- Yellow guardrails are installed around the top section of the machinery, indicating safety measures to prevent falls during operation.
- Additional yellow handrail sections can also be seen on other parts of the structure, ensuring operator protection while working at height.

3. **Identification Markings**:
- "STR" is prominently displayed on the front panel of the machine, possibly denoting the manufacturer's name or model identifier.
- Below "STR," there's another label reading "SB168," which could represent a specific product line or serial number associated with the equipment.

4. **Covered Section**:
- At the bottom right corner of the image, there is a covered portion of the equipment. The cover bears white text that reads "SB168." This indicates either a storage compartment or a protective casing over sensitive internal mechanisms.

5. **Surrounding Environment**:
- In the background, various pieces of industrial equipment and containers are visible, suggesting a typical setting such as a warehouse, factory yard, or construction site.
- Trees and greenery indicate that the location may not be entirely urban but rather semi-rural or suburban.

6. **Sky Conditions**:
- Above, the sky is mostly clear with scattered clouds, implying fair weather conditions suitable for outdoor operations.

Analysis:
The described object is most likely a type of pump station designed for heavy-duty use, probably utilized in industries like construction, mining, or wastewater management due to its robust build and extensive piping setup. Pumps labeled with identifiers like SB168 often denote standardized models within these sectors, known for their reliability and durability. Given the context clues provided by the surrounding environment (industrial equipment, trees), we infer that this equipment plays a critical role in managing fluid flow efficiently across different sites where liquid transfer is necessary.

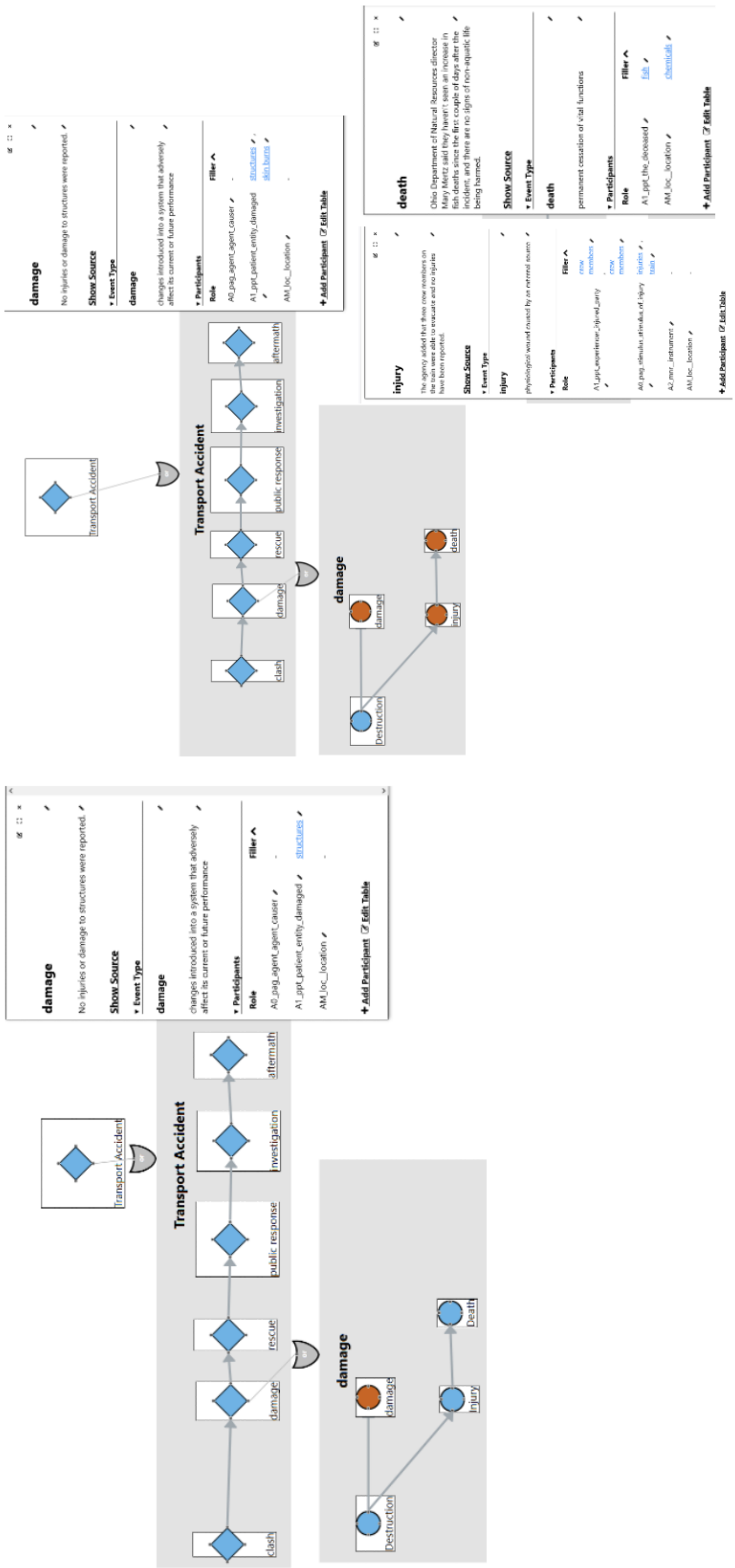

Figure 5: Same“transport accident” schema with “damage” events extracted from earliest date information and then with additional extracted information up to 10 days later

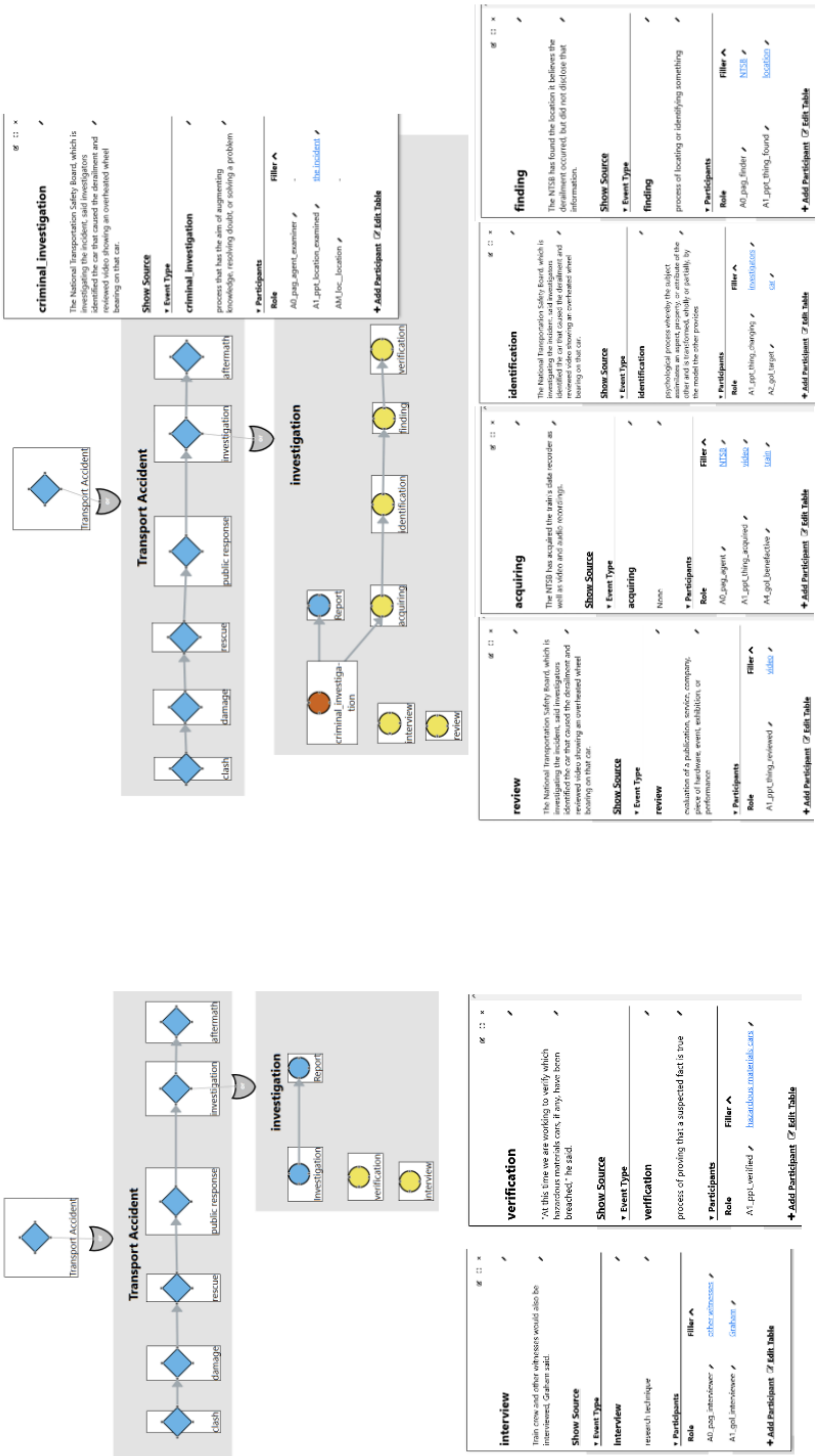

Figure 6: Same“transport accident” schema with “investigation” events extracted from earliest date information and then with additional extracted information up to 10 days later

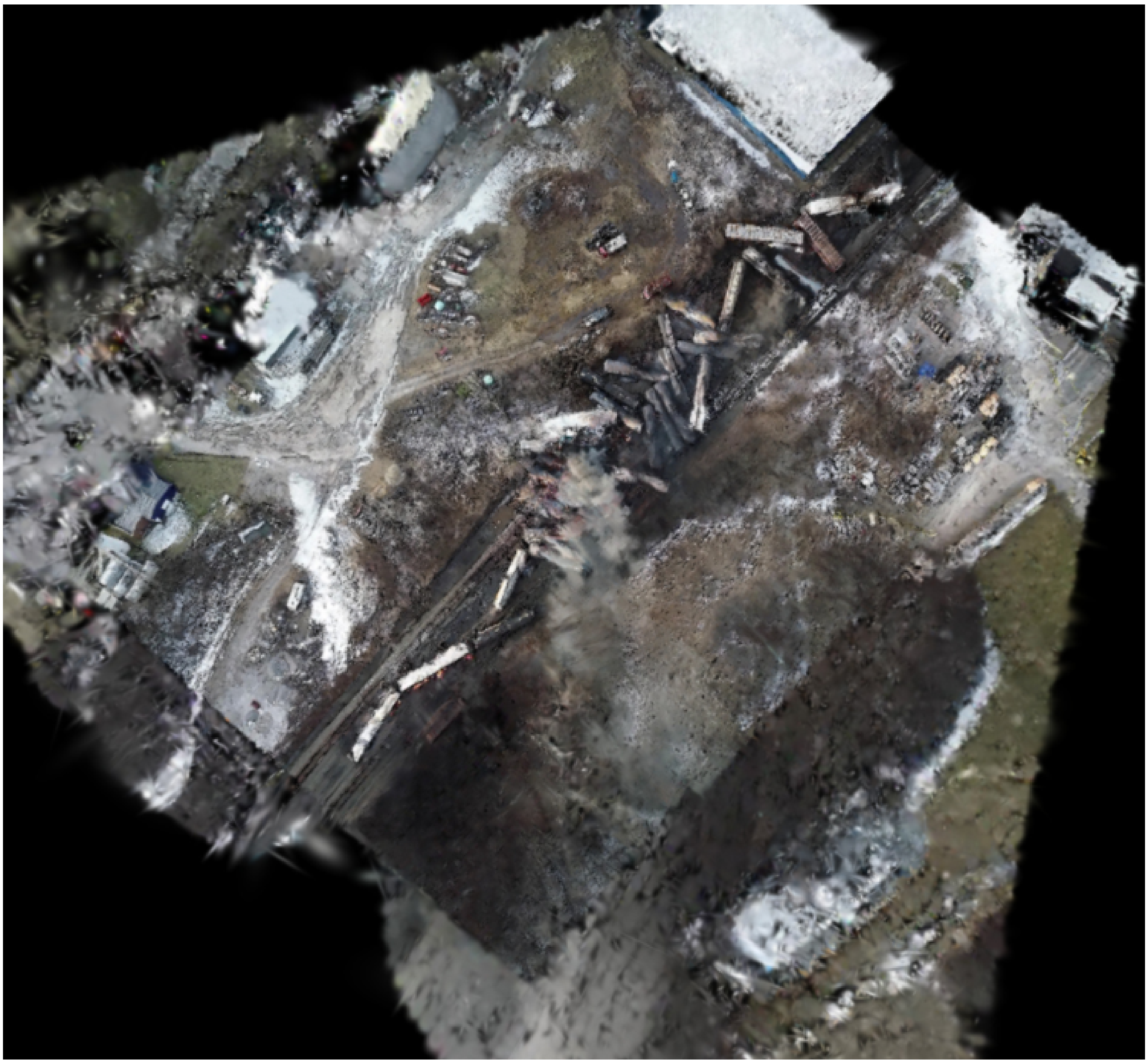

Figure 7: Flattened view of the 3D simulation