Towards the Enrichment of Conceptual Models with Multimodal Data

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

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

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

Abstract

Conceptual models are essential for designing, analyzing, and communicating complex systems. However, traditional modeling languages such as UML, BPMN, and ArchiMate primarily rely on textual and symbolic representations, which can limit their expressiveness and accessibility, especially for non-expert stakeholders. To address this challenge, we introduce a framework for *Multimodal-Enriched Conceptual Modeling* (MMeCM) that integrates videos, images, and audio directly into model elements. Our approach enables modelers to attach contextual multimedia references to processes, entities, and relationships, effectively grounding abstract concepts in tangible real-world artifacts. We make three key contributions: (1) a quantitative analysis of concept enrichability using the OntoUML/UFO Catalog, identifying which elements benefit from multimodal representation; (2) the design and implementation of a generalizable framework for embedding multimodal data across different modeling languages; and (3) a qualitative user study, grounded in the *Technology Acceptance Model*, evaluating the perceived usefulness and usability of multimodal-enriched models, together with a dataset of more than 12K multimodal-enriched natural language elements found in conceptual models. Our evaluation shows that a majority of natural language elements in conceptual models can be effectively augmented with multimedia, and user feedback indicates a strong positive reception of MMeCM.

Keywords: Conceptual Modeling, Multimodal Data, Model Enrichment, Knowledge Representation, Interactive Models

1. Introduction

Modeling languages like the Business Process Model and Notation (BPMN) and Unified Modeling Language (UML) have been developed to accommodate different perspectives and granularities [21]. Such languages, combined with meta-modeling and analytics techniques, help enterprises adapt to evolving requirements by forming the backbone of strategic planning and digital transformation initiatives [14]. Yet, most conceptual models still rely on textual or symbolic representations, which may limit their accessibility [23] and expressiveness [4], especially when dealing with real-world contexts replete with multimedia evidence [12]. Efforts to use prompting strategies with large language models [5, 19, 22] offer new possibilities for automating or augmenting modeling tasks, yet most approaches do not systematically incorporate multimodal data such as videos and images.

Our previous work has contributed to bridging this gap by augmenting process models with extended reality and multimodal evidence [9]. One line of research focuses on discovering and

annotating event logs with contextual videos and images [10, 12], while another explores tailoring these artifacts to specific stakeholder jargon [13], and even interpreting diagrammatic notations through multimodal references [11]. These approaches align with broader efforts to improve annotation tools for process information extraction [20] and harness large language models for advanced process tasks [19]. Yet, the question remains how to fully integrate the wealth of multimodal data sources–videos, audio, sensor streams–into conceptual models without overcomplicating them and increasing the cognitive load. Recent advancements in multimedia technologies [15, 18] have opened the door for a novel form of conceptual model enrichment. By integrating videos, images, and audio clips directly into modeling elements, it becomes possible to provide supplementary material that can significantly increase the clarity and fidelity of models—helping *to bridge the gap between the abstract representations found in models and the physical, tangible artifacts or scenarios they represent*. In this paper, we propose and evaluate a new approach to enrich conceptual modeling with multimodal data. This paper focuses on the following three research objectives (ROs):

RO1: Analyze curated, high-quality conceptual models to understand which natural language concepts can be clearly represented using multimodal data (such as images or audio), and which concepts are too abstract and are better conveyed through symbolic representations, like icons.

RO2: Outline and implement a generalizable framework for integrating multimodal data across different modeling languages, followed by an evaluation of this framework using the high-quality curated conceptual models as a testbed.

RO3: Conduct a qualitative assessment based on the *Technology Acceptance Model* to investigate how users would perceive, accept, and potentially adopt multimodal-enriched conceptual models.

The remainder of this paper is structured as follows. Section 2 outlines our research methodology and an illustrative example. Section 3 presents our proposed framework for integrating multimodal data into conceptual models. In Section 4, we discuss the evaluation of our approach. Section 5 reviews related work. Finally, Section 6 concludes the paper. The supplementary material[1] contains artifacts, evaluations, and implementation.

2. Methodology: The General Perspective

In this section, we detail the applied research methodology for enriching conceptual models with multimodal data. The methodology is structured into several subsections that build on one another: we first formalize the structure of conceptual modeling languages, then provide a general method for multimodal enrichment of conceptual models, and finally discuss its specific instantiation for the OntoUML modeling language as a running case.

Figure 1 presents an overview of the methodology for integrating multimodal data into conceptual models. The process spans three interconnected spaces: the *Conceptual Model Space*, the *Embedding Space*, and the *Multimodal Data Space*. The conceptual model consists of structured elements, such as kinds, roles, qualities, and relations, which are linked through mediation, characterization, and inherence relationships. These elements are projected into an embedding space, where they are transformed into distributed vector representations. Simultaneously, multimodal data—such as images and audio—are encoded into corresponding embeddings. The embedding space serves as an intermediary, aligning the conceptual structures with multimodal data representations, thereby enabling enriched semantic connections.

[1] Supplementary material: `https://github.com/aleksandargavric/mm_cm_enrichment`.

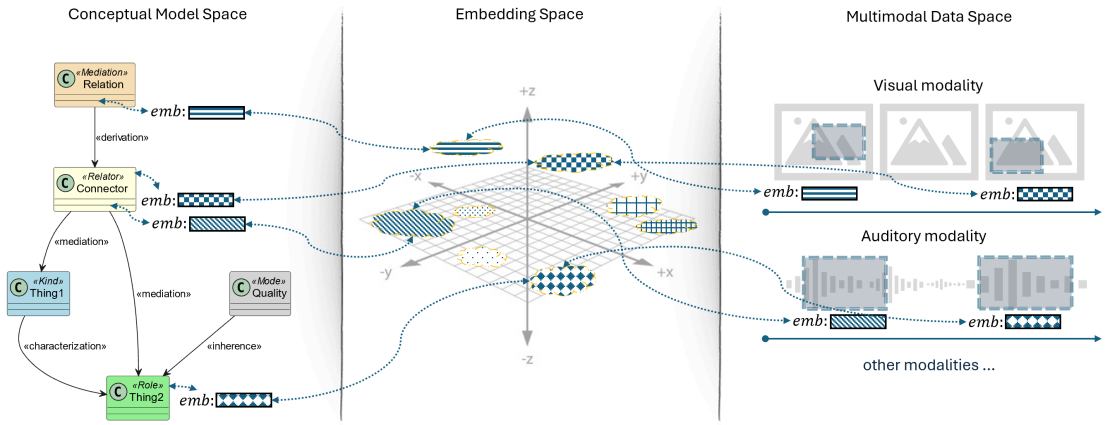

Fig. 1. Overview of the methodology for enriching conceptual models with multimodal data. Patterned regions illustrate embeddings. An embedding space is a continuous, high-dimensional vector space in which discrete data are represented as vectors, such that the spatial relationships among the vectors reflect meaningful similarities or patterns in the original data.

Preliminaries.

Let L be a set of conceptual modeling languages. For any language $\ell \in L$, its metamodel is defined as $\mathcal{M}_\ell = (E_\ell, R_\ell, \alpha_\ell)$, where E_ℓ denotes the set of modeling elements (e.g., classes, entities, activities), R_ℓ denotes the set of relationships (e.g., associations, generalizations), and α_ℓ is the set of syntactic and semantic constraints. A concrete model M_ℓ conforming to $\mathcal{M}_\ell$ is represented as $M_\ell = (E_\ell^*, R_\ell^*, \alpha_\ell^*)$, with $E_\ell^* \subseteq E_\ell$, $R_\ell^* \subseteq R_\ell$, and $\alpha_\ell^* \subseteq \alpha_\ell$.

We introduce the concept of multimodal enrichment by extending the conventional model structure.

Definition 2.1 (Multimodal Data) *Let Γ be a finite set of multimedia types, where $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_m\}$ and each γ_j represents a distinct media type (e.g., image, audio, video).*

Definition 2.2 (Attachment Function) *For a given conceptual model $M_\ell = (E_\ell^*, R_\ell^*, \alpha_\ell^*)$, define the attachment function $\mathrm{Att} : E_\ell^* \to 2^\Gamma$ such that for every element $e \in E_\ell^*$, $\mathrm{Att}(e) \subseteq \Gamma$.*

Definition 2.3 (Multimodal-Enriched Conceptual Model) *A multimodal-enriched conceptual model is defined as $M_\ell^+ = (E_\ell^*, R_\ell^*, \alpha_\ell^*, \mathrm{Att})$, where Att is the attachment function as specified above.*

To instantiate this methodology, we outline the following steps: (1) Extend existing modeling tools to allow each element $e \in E_\ell^*$ to carry an attribute $\mathrm{Att}(e)$ that references external multimedia resources. (2) Define a metadata function $\mu : \Gamma \to \mathcal{M}$, where for any $\gamma \in \Gamma$, $\mu(\gamma)$ provides essential metadata such as the URI, file format, and checksum. (3) Ensure that the enriched model M_ℓ^+ adheres to the original constraints α_ℓ by validating $\mathrm{Att}(e)$ for all $e \in E_\ell^*$.

2.1. An Illustrative Scenario

To demonstrate the practical application of our methodology, we present an illustrative scenario focused on the furniture assembly process. This scenario provides an in-depth view of how conceptual models can be enriched with multimodal data, including visual and auditory modalities, enabling enhanced semantic representation and reasoning.

Fig. 2 showcases a structured conceptual model of a furniture assembly process, integrating multiple modalities into an enriched representation. The model consists of key entities such

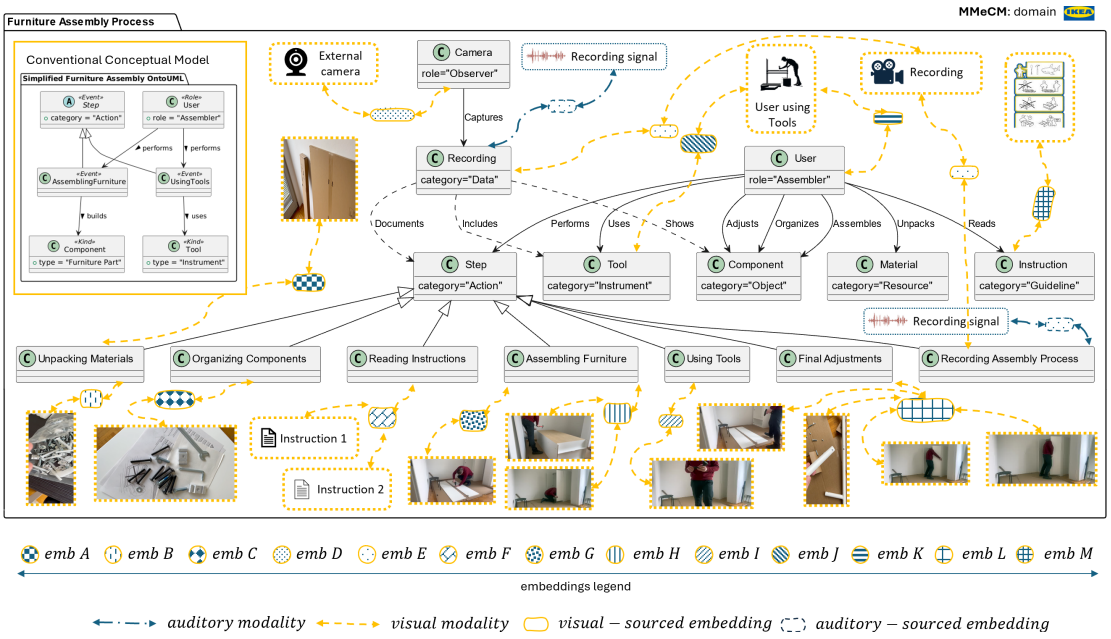

Fig. 2. Illustrative scenario: Multimodal enrichment of the furniture assembly process. *(If labels are not readable in print, focus on the layout and structure—refer to the supplementary material for a high-resolution version.)*

as `User`, `Step`, `Tool`, `Component`, and `Material`, each categorized under specific conceptual roles (e.g., `Action`, `Instrument`, `Object`, `Resource`). These elements are interconnected through relations, such as `performs`, `uses`, and `adjusts`, forming the structural foundation of the assembly process.

The conceptual model elements are projected into an embedding space, where each entity e_i is mapped to a corresponding vector representation $\mathbf{v}_i$. In particular, let $\mathcal{E}$ denote the set of entities, and let $f : \mathcal{E} \to \mathbb{R}^d$ be the embedding function mapping each entity e_i to a d-dimensional space: $\mathbf{v}_i = f(e_i), \quad \forall e_i \in \mathcal{E}$. This embedding allows semantic comparisons and proximity-based retrieval of related concepts.

Multimodal data sources, including visual and auditory signals, are aligned with the conceptual model through their respective embeddings. The embeddings from images (`visual modality`) and audio recordings (`auditory modality`) are extracted via pretrained model, concretely [15]: $\mathbf{v}_j^{(vis)} = g_{\text{vision}}(I_j), \quad \mathbf{v}_k^{(aud)} = g_{\text{audio}}(A_k)$, where I_j and A_k denote image and audio samples, and $g_{\text{vision}}, g_{\text{audio}}$ are modality-specific embedding functions.

Each step in the assembly process—such as `Unpacking Materials`—is documented through multimodal evidence. The model captures visual cues (e.g., images of screws, planks, and assembly instructions) and auditory recordings (e.g., the sound of a screwdriver or user commentary). The embeddings are dynamically updated based on new observations: $\mathbf{v}_t^{(step)} = \alpha \mathbf{v}_t^{(step)} + (1 - \alpha)\mathbf{v}_t^{(\text{new})}$, where $\mathbf{v}_t^{(step)}$ represents the embedding of an assembly step at time t, and $\mathbf{v}_t^{(\text{new})}$ is the new multimodal update. The weighting factor α determines the balance between prior knowledge and new data.

By integrating multimodal embeddings, the model enables richer semantic reasoning. For example, the action `Using Tools` is associated with a set of tools (e.g., screwdriver, hammer), each linked to corresponding embeddings from prior tasks. Given a new tool T_x, similarity-based reasoning can determine its function by computing its distance to known tool embeddings. If sim exceeds a predefined threshold, the model can infer the tool's function and suggest appropriate usage instructions.

3. MMeCM: A Framework for Integrating Multimodal Data into Conceptual Modeling

In this section, we present our proposed framework for integrating multimodal data into conceptual models, producing *Multimodal-enriched Conceptual Models* (**MMeCM**). Our framework consists of three primary stages: (1) extraction of all natural language elements from conceptual models while preserving their contextual links, (2) computation of multimodal embeddings for each extracted element, and (3) performing similarity matching between these embeddings and multimodal data. For a detailed implementation of the MMeCM framework, we invite readers to consult the supplementary materials accompanying this work[1].

In our approach, the similarity matching is facilitated by ImageBind [15], a method that learns a joint embedding space across six modalities (in particular, images, text, audio, depth, thermal, and inertial measurement units such as accelerometer and gyroscope). We design our similarity matching to act as the enrichment step by *linking any of these six modalities to the corresponding natural language elements in conceptual models*.

As part of our framework evaluation and to facilitate further research in multimodal conceptual modeling, we contribute the *Multimodal-enriched dataset of conceptual models*, a curated dataset comprising natural language elements extracted from the OntoUML models. This dataset $\mathcal{D} = \bigcup_{i=1}^{15} \mathcal{D}_i$ denotes the complete dataset, where each chunk $\mathcal{D}_i = \{(n_j, v_j) \mid n_j \in \mathcal{N}_i, v_j = F(n_j)\}$ consists of a finite set of natural language elements n_j and their corresponding multimodal embeddings $v_j \in \mathbb{R}^d$ computed by the embedding function F. Each n_j originates from a conceptual model element $e_j \in E$, preserved through a mapping $\mu : \bigcup_i \mathcal{N}_i \to E$. The embeddings reside in a shared multimodal space $\mathbb{R}^d$ such that a similarity function $\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (e.g., cosine similarity) enables cross-modal comparisons between textual inputs and elements from other modalities $\Gamma = \{\text{image}, \text{text}, \text{audio}, \text{depth}, \text{thermal}, \text{IMU}\}$. This dataset supports multiple downstream tasks, such as multimodal retrieval ($\arg\max_{x \in \mathcal{D}} \sigma(v_q, v_x)$ for a query v_q from any modality), semantic clustering over $\mathcal{D}$ using embedding topology, or evaluation of alignment functions $\phi : \Gamma \to \mathcal{N}$ in AI-assisted modeling scenarios.

Prototype Walkthrough.

In our prototype framework, we (1) extract each natural language term from the model, (2) clean and normalize the extracted term, (3) convert each term into a vector via the multimodal encoder, and (4) compute similarity scores between vectors to attach matching multimodal data.

3.1. Implementation

Let a conceptual model be denoted as M, which contains a set of elements E and their associated natural language descriptions. The goal is to extract each natural language description while keeping track of its original position in M. Let $\mathcal{N}$ be the set of all extracted natural language values, and let $\mu : \mathcal{N} \to E$ be the mapping that associates each natural language description with its corresponding model element. After extraction, each $n \in \mathcal{N}$ is transformed into an embedding $v \in \mathbb{R}^d$ using a multimodal model. The similarity function $\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ then computes the affinity between any two embeddings, thereby enabling the enrichment with modalities from the set Γ.

Extraction of Natural Language Elements.

To initiate the enrichment process, we extract all natural language elements from the conceptual model while preserving their contextual links. This step lays the foundation for linking the extracted textual descriptions with the corresponding multimodal data. Alg. 1 outlines the extraction of these elements from the model.

Algorithm 1 Extract Natural Language Elements from Conceptual Models

Input: Conceptual model M with element set E
Output: Set of natural language values $\mathcal{N}$ and mapping $\mu : \mathcal{N} \to E$
Initialize $\mathcal{N} \leftarrow \emptyset$ and mapping $\mu \leftarrow \{\}$
for each element $e \in E$ **do**
    Identify natural language properties $P(e)$ associated with e
    for each property $p \in P(e)$ **do**
        Extract natural language value n from p
        $\mathcal{N} \leftarrow \mathcal{N} \cup \{n\}$
        Update mapping $\mu(n) \leftarrow e$
    end for
end for
return $(\mathcal{N}, \mu)$

Following the extraction, the set of natural language values $\mathcal{N}$ and the mapping μ serve as the basis for the *Embedding and Similarity Matching* stage in our pipeline.

Embedding and Similarity Matching for Enrichment.

The framework proceeds to compute multimodal embeddings for each natural language element. This embedding step leverages a multimodal model to transform natural language into a joint embedding space, where similarity matching can be performed across different data modalities. Alg. 2 illustrates how the textual data is preprocessed, embedded, and subsequently aligned with multimodal data via similarity scoring.

Algorithm 2 *Multimodal-enriched Conceptual Models* (**MMeCM**)

Input: Set of natural language values $\mathcal{N}$ (produced with Alg. 1), multimodal model F, and modality set Γ, batch size B, and conceptual model M
Output: Embedding mapping $\eta : \mathcal{N} \to \mathbb{R}^d$ and similarity matrices for enrichment
Partition $\mathcal{N}$ into batches $\{N_1, N_2, \ldots, N_k\}$ of size B
for each batch N_i **do**
    Prepare a list T containing all values in N_i
    Transform T using the natural language preprocessing function $T' = \mathcal{T}(T)$, in the context of conceptual model M
    Compute embeddings: $E = F(T')$, where $E \subset \mathbb{R}^d$
    for each embedding $e_j \in E$ associated with text $t_j \in T$ **do**
        Set $\eta(t_j) \leftarrow e_j$
    end for
    Compute similarity scores matrix S such that for any pair (t_j, t_k),

$$S_{jk} = \sigma(e_j, e_k) = \text{softmax}(e_j^\top e_k)$$

    Store S and η for batch N_i for further matching with data from any modality in Γ
end for
return η and all computed similarity matrices

Computation of embeddings using a multimodal model, followed by the construction of similarity matrices, facilitate matching between the natural language embeddings and the various modalities in Γ, thus enabling the multimodal enrichment of conceptual models through a flexible approach that can be adapted to diverse conceptual modeling languages and extended to integrate additional modalities as required.

3.2. Dataset Contribution

To facilitate further research in multimodal conceptual modeling, we contribute the *Multimodal-enriched dataset*, a curated dataset comprising natural language elements extracted from existing conceptual models in the OntoUML/UFO catalogue [1]. The dataset contains 12,300 unique entries that are structured into 15 chunks to support modular experimentation. Each entry corresponds to a filtered (i.e., exact duplicates removed) natural language element linked to a specific OntoUML/UFO modeling construct, maintaining the semantic integrity and contextual role (through preserved model attributes) of the original model element, enriched with a *multimodal embedding* computed using ImageBind [15]. This collection reflects a diverse sample of modeling scenarios and linguistic formulations that appear in OntoUML practices. Each embedding resides in a *joint latent space* that supports *comparisons across six modalities*: text, image, audio, depth, thermal, and IMU. The dataset thereby enables similarity queries between natural language model elements and arbitrary modality inputs using standard distance metrics such as cosine or Euclidean similarity. The resulting representation preserves conceptual modeling semantics for novel cross-modal retrieval tasks and enrichments, laying the foundation for multimodal reasoning, alignment, and AI-assisted modeling support tools.

4. Evaluation

In this section, we discuss the scalability, performance, cost of operation, security and privacy concerns, as well as the systematic evaluation of our methodology and framework through both quantitative and qualitative analyses. We first provide an empirical analysis of the dataset and retrieval performance through similarity matrices. Then, we conduct a user study to assess the practical usability and acceptance of our approach using the *Technology Acceptance Model* (TAM) [6]. Finally, we outline key findings, limitations, and threats to validity.

Scalability Evaluation.

Our framework is designed with modularity and extensibility in mind, ensuring that it scales well with increasing dataset sizes and additional modalities. Thanks to the use of chunked processing and embedding caching mechanisms, performance remains robust even when operating over tens of thousands of elements. During experimentation, we observed linear growth in both computation time and memory consumption relative to dataset size, which confirms predictable scalability for future expansions.

Performance Evaluation.

One of the key strengths of our approach lies in its ability to deliver real-time performance. All computations, including embedding generation, similarity calculations, and retrieval tasks, are executed locally without reliance on cloud-based services and *APIs*. This architecture ensures minimal latency and allows for instantaneous feedback, making the solution viable for interactive use cases and deployments in latency-sensitive environments such as AR/XR applications.

Cost of Operation.

By eliminating the need for external servers, cloud subscriptions, or internet connectivity during operation, our solution significantly reduces the cost of deployment and maintenance. The entire pipeline can be run efficiently on a modern consumer-grade laptop with GPU acceleration (we used *NVIDIA A40 GPU* with 48 GB GDDR6 ECC Memory), making it accessible for academic, industrial, or personal use. This low cost of ownership broadens the framework's applicability across institutions with limited budgets.

Security and Privacy Concerns.

Security and privacy were central considerations in our system design. Since all processing is performed locally, no sensitive data is transmitted over the internet or stored in third-party services. This architecture inherently mitigates risks associated with data leakage or unauthorized access, ensuring full compliance with privacy regulations such as GDPR (The General Data Protection Regulation, EU-2016/679). Furthermore, local execution allows users full control over their data, aligning with privacy-by-design principles.

4.1. Quantitative Evaluation

We have structured our created dataset into 15 distinct chunks, each containing extracted and filtered natural language elements (with exact duplicates removed) along with their corresponding multimodal embeddings. The dataset distribution across chunks is as follows: 1265, 702, 999, 1562, 571, 612, 691, 821, 627, 499, 981, 458, 843, and 669 elements per chunk, each corresponding to the number of unique natural language elements found in a 10 model subsets of OntoUML/UFO Catalogue. As these embeddings allow for similarity comparisons across six modalities using standard distance metrics such as Euclidean and cosine distance, we provide in our supplementary material similarity matrices (per chunk, as there are 12,300 entries in total) where the diagonal represents self-similarity (value of 1), demonstrating consistency in embedding alignment.

To assess the relevance of multimodal retrievals in real-world scenarios, we consider three evaluation approaches: *(A) domain expert annotations*, where experts review real-life conceptual model examples and mark potential spots where multimodal data could enrich the model; *(B) large-scale pretrained language model annotation*, taking foundational pretrained large language model to simulate the annotation process based on approach (A), and finally, *(C) data-driven retrieval assessment*, by collecting multimodal data samples and evaluating the retrieval quality directly. We opt for a combination of (A) and (B). A golden dataset is created through domain expert annotations identifying multimodal enrichment spots in real-life conceptual models (through approach A). Since human annotation is time-consuming and susceptible to lapsus errors, we complement it with a large language model (LLM) generated dataset trained using golden dataset examples and annotation instructions (which is approach B, yet augmented with approach A result). For the expert annotation process, we developed an annotation tool as shown in Figure 3. The tool allows annotators to mark and revise elements within conceptual models (approach A), specifying whether visual or auditory enrichment is relevant. Its initial suggestions are produced using an open-weight LLM [16] that offers advanced reasoning capabilities in a more resource-efficient package (approach B). To generate automated annotations, we use the prompt given in the supplementary material.

4.2. Qualitative Evaluation

To assess the practical usability of our approach, we conducted a user study grounded in the Technology Acceptance Model (TAM). Originally developed by [6], TAM is a widely adopted framework in system adoption research, that posits that a user's acceptance of a technology is primarily influenced by two core factors: *Perceived Usefulness* (PU) and *Perceived Ease of Use* (PEOU). These, in turn, affect the user's *Attitude Toward Using the system* (ATT), which shapes their *Behavioral Intention to Use* (BI), ultimately leading to *Actual Use*.

In our study, we applied this structure to evaluate our approach: Perceived Usefulness (PU) — "Does MMeCM enhance your conceptual modeling tasks?"; Perceived Ease of Use (PEOU) — "Is MMeCM intuitive and user-friendly?"; Attitude Toward Using (ATT) — "Would you consider using MMeCM in your workflow?"; Behavioral Intention (BI) — "Would you recommend MMeCM to colleagues?". We engaged a diverse group of participants with varying levels

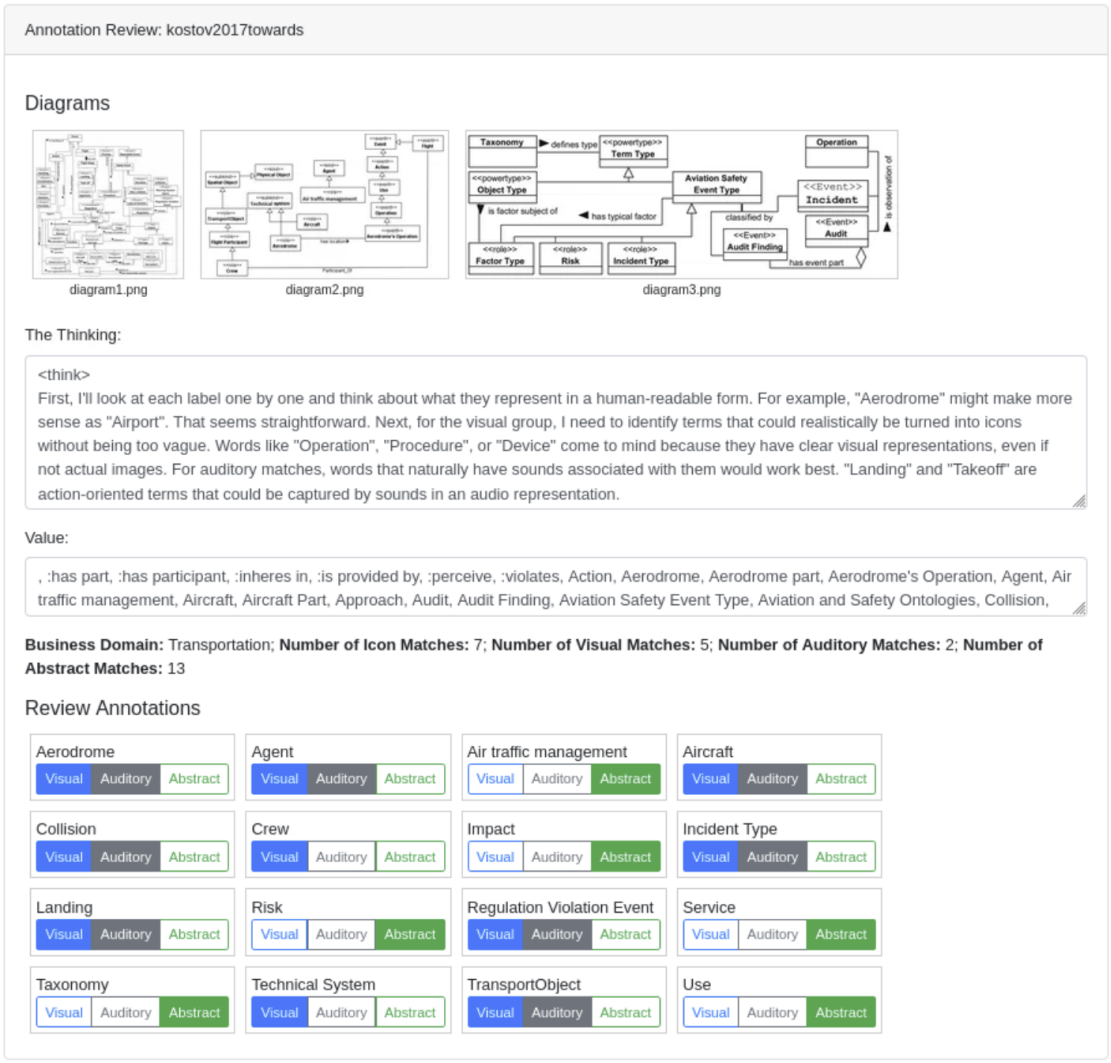

Fig. 3. Annotation tool used for expert-based multimodal enrichment identification. Its initial suggestions are produced using an LLM as described. The OntoUML that is visualized in the example is [17].

of expertise in conceptual modeling: (A) two PhD holders in conceptual modeling, (B) three PhD candidates with over two years of experience in conceptual modeling, and (C) five students with foundational knowledge of conceptual modeling. Participants responded using a 5-point Likert scale.

4.3. Discussion: Key Findings, Limitations, and Threats to Validity

Our evaluation yielded several key findings. Quantitatively, the multimodal embeddings demonstrated *consistent alignment in a multimodal embedding space*, as evidenced by the similarity matrices, validating the structural integrity of our dataset and retrieval mechanism. The combined annotation approach—leveraging domain experts and a large language model—showed that more than 68% of *natural language elements in conceptual models can be enriched with multimodal data*. Qualitatively, results from the TAM-based user study revealed a positive reception (4.2/5.0) toward MMeCM, particularly in terms of perceived usefulness and ease of use, with participants highlighting MMeCM's potential to improve expressiveness and accessibility in conceptual modeling tasks.

Nonetheless, several limitations and threats to validity remain. First, the dataset is built on a specific modeling catalog (OntoUML/UFO), which may limit generalizability to other modeling paradigms. Second, while the LLM-assisted annotation complements expert input, it may introduce subtle biases based on prompt phrasing and training data. Third, the user study was conducted with a relatively small and academically-inclined sample, which might not fully represent practitioners in industry settings. Lastly, while MMeCM suggests multimodal enrichments, the evaluation does not yet measure downstream effects on task performance or decision-making quality. Further evaluation will address these concerns by expanding domain coverage, increasing participant diversity, and incorporating longitudinal usage studies.

5. Related Work

A variety of research efforts explore the intersection of conceptual modeling and advanced data-driven techniques, including machine learning, and multimodal analysis. One notable body of work uses few-shot prompt learning, which integrate large language models for automating or assisting modeling tasks, also point to new opportunities for model completion and augmentation [5].

Beyond AI integration, information visualization and human-computer interaction in conceptual modeling have gained significant attention. One taxonomy details a spectrum of visualization and interaction techniques applicable to modeling environments, illustrating how advanced visual concepts might enhance the creation and inspection of models [3]. This aligns with ongoing work in meta-modeling platforms, where frameworks such as CMAG propose ways to incorporate generative AI into conceptual modeling [8], and extensions to spatial conceptual modeling investigate how physical locations and augmented reality can be tied to meta-modeling primitives [7]. Several recent efforts highlighted the promise of enriching event logs and business process (conceptual) elements with multimodal data for use in process mining task [10, 12]. Related work focuses on stakeholder-specific representations of such data [13] and explores AI-driven interpretation of UML diagrams in a multimodal context [11]. Tools for efficient annotation and extraction of process information further support these new frontiers [20]. Taken together, these bodies of work motivate the need for a systematic methodology and infrastructure that can unify textual, symbolic, and multimodal data streams in conceptual modeling environments.

While modeling languages (i.e. *UML*, *BPMN*, and *ArchiMate*) provide extension mechanisms—such as stereotypes and profiles—used in tools like *BizAgi*, *Visual Paradigm*, and *Bizzdesign* to attach multimedia elements, these approaches typically treat such data as abstract (icon) annotations. The model engineering tradition to express the models and its instances (e.g., [2]), improve modeling flexibility but fall short in semantically integrating the actual instances of multimodal data. Our approach goes further by directly linking abstract concepts to instance-level multimodal data, enabling structured interpretation and interaction beyond what conventional extensions afford.

6. Conclusion

This paper presents a novel approach to enriching conceptual models through the integration of multimodal data, enhancing their expressiveness and accessibility. By bridging symbolic abstractions with real-world references, our MMeCM framework demonstrates both technical feasibility and user acceptance. The results highlight a strong potential for multimodal augmentation in conceptual modeling, with over two-thirds of natural language elements found suitable for such enrichment. We presented a structured roadmap for advancing a multimodal and collaborative conceptual modeling framework, emphasizing both technical enhancements and user-centered design. By prioritizing tasks such as modality expansion, tool integration,

federated model support, and inclusive usability testing, we aim to foster a more expressive, scalable, and accessible modeling environment.

References

[1] Barcelos, P.P.F., Sales, T.P., Fumagalli, M., Fonseca, C.M., Sousa, I.V., Romanenko, E., Kritz, J., Guizzardi, G.: A fair model catalog for ontology-driven conceptual modeling research. In: International Conference on Conceptual Modeling. pp. 3–17. Springer (2022)

[2] Bézivin, J.: 2 building modeling languages. Domain-Specific Languages: Effective Modeling, Automation, and Reuse p. 25 (2023)

[3] Bork, D., De Carlo, G.: An extended taxonomy of advanced information visualization and interaction in conceptual modeling. Data & Knowledge Engineering 147, pp. 102209 (2023)

[4] Bork, D., Schrüffer, C., Karagiannis, D.: Intuitive understanding of domain-specific modeling languages: Proposition and application of an evaluation technique. In: Laender, A.H.F., Pernici, B., Lim, E., de Oliveira, J.P.M. (eds.) Conceptual Modeling - 38th International Conference, ER 2019, Salvador, Brazil, November 4-7, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11788, pp. 311–319. Springer (2019)

[5] Chaaben, M.B., Burgueño, L., Sahraoui, H.: Towards using few-shot prompt learning for automating model completion. In: 2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER). pp. 7–12 (2023)

[6] Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly pp. 319–340 (1989)

[7] Fill, H.G.: Spatial Conceptual Modeling: Anchoring Knowledge in the Real World, pp. 35–50. Springer Nature Switzerland, Cham (2024)

[8] Fill, H.G., Härer, F., Vasic, I., Borcard, D., Reitemeyer, B., Muff, F., Curty, S., Bühlmann, M.: Cmag: A framework for conceptual model augmented generative artificial intelligence (12 2024)

[9] Gavric, A.: Enhancing process understanding through multimodal data analysis and extended reality. In: Companion Proceedings of the 16th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling and the 13th Enterprise Design and Engineering Working Conference (November 28 - December 01 2023)

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

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

[12] Gavric, A., Bork, D., Proper, H.: Multimodal process mining. In: 26th International Conference on Business Informatics (CBI) (2024)

[13] Gavric, A., Bork, D., Proper, H.: Stakeholder-specific jargon-based representation of multimodal data within business process. In: Companion Proceedings of the 17th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling (PoEM Forum 2024) (2024)

[14] Gils, B.v., Proper, H.A.: Next-Generation Enterprise Modeling, pp. 279–305. Springer Nature Switzerland, Cham (2023)

[15] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)

[16] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

[17] Kostov, B., Ahmad, J., Křemen, P.: Towards Ontology-Based Safety Information Management in the Aviation Industry. In: Lecture Notes in Computer Science, vol. 10034 LNCS, pp. 242–251 (2017)

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

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

[20] Neuberger, J., Herrmann, J., Käppel, M., van der Aa, H., Jablonski, S.: Teapie: A tool for efficient annotation of process information extraction data. In: International Conference on Cooperative Information Systems. pp. 405–410. Springer (2024)

[21] Proper, H.A., van Gils, B., Haki, K.: Final Conclusions and Outlook, pp. 311–314. Springer Nature Switzerland, Cham (2023)

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

[23] Sarioglu, A., Metin, H., Bork, D.: Accessibility in conceptual modeling - A systematic literature review, a keyboard-only UML modeling tool, and a research roadmap. Data Knowl. Eng. 158, pp. 102423 (2025)