

Moving Pictures of Thought: Extracting Visual Knowledge in Charles S. Peirce's Manuscripts with Vision-Language Models

Carlo Teo Pedretti¹ , Davide Picca² , and Dario Rodighiero³ 

¹ Department of Classics, University Sapienza of Rome, Rome, Italy

² Department of Language and Communication Sciences, University of Lausanne, Lausanne, Switzerland

³ Campus Fryslân, University of Groningen, Groningen, The Netherlands

Abstract

Diagrams are crucial yet underexplored tools in many disciplines, demonstrating the close connection between visual representation and scholarly reasoning. However, their iconic form poses obstacles to visual studies, intermedial analysis, and text-based digital workflows. In particular, Charles S. Peirce consistently advocated the use of diagrams as essential for reasoning and explanation. His manuscripts, often combining textual content with complex visual artifacts, provide a challenging case for studying documents involving heterogeneous materials. In this preliminary study, we investigate whether Visual Language Models (VLMs) can effectively help us identify and interpret such hybrid pages in context. First, we propose a workflow that (i) segments manuscript page layouts, (ii) reconnects each segment to IIIF-compliant annotations, and (iii) submits fragments containing diagrams to a VLM. In addition, by adopting Peirce's semiotic framework, we designed prompts to extract key knowledge about diagrams and produce concise captions. Finally, we integrated these captions into knowledge graphs, enabling structured representations of diagrammatic content within composite sources.

Keywords: visual language models, diagrams, IIIF, semantic web, visual semiotics

1 Introduction

Diagrams play a central role in many forms of reasoning, from mathematics to philosophy and religious art [12; 20; 29]. Among the most prominent theorists of diagrammatic reasoning is Charles S. Peirce, who conceived of diagrams as a subtype of icon capable of representing and manipulating internal structures through visual means [31; 32]. In his unpublished manuscripts, diagrams such as *Existential Graphs* illustrate logical inferences via spatial configurations, offering a visual alternative to symbolic logic. Peirce referred to these constructs as “moving pictures of thought” [24] (*Collected Papers* 4.8), underlining their dynamic and epistemological function.

This idea finds material expression in Peirce's manuscripts, where textual content, visual artifacts, and complex layouts are seamlessly integrated [18]. These documents reflect both his theoretical commitment to diagrammatic reasoning and its practical development through layered and visually structured writing. However, this visual richness remains largely inaccessible in existing printed editions, which are compiled under severe editorial constraints [16; 17].

Building on recent advances in the textual analysis of Peirce's manuscript *Prolegomena to an Apology for Pragmaticism* (PAP) [25], we extend the investigation to his visual thinking. This exploratory study investigates the extent to which Vision Language Models (VLMs) can engage with

Carlo Teo Pedretti, Davide Picca, and Dario Rodighiero. “Moving Pictures of Thought: Extracting Visual Knowledge in Charles S. Peirce's Manuscripts with Vision-Language Models.” In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1434–1447. <https://doi.org/10.63744/fkFGJ6wSzDPV>.

© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

diagrams as operative semiotic forms. To address this question, we propose a flexible and interoperable workflow for extracting structured knowledge from multimodal documents that supports integration within Linked Open Data (LOD) environments. Our modular workflow begins with the segmentation of page layouts to isolate diagrams, which are then linked IIF annotations. These fragments are submitted to a VLM via prompt-based interactions informed by Peirce’s semiotic theory, with the aim of generating structured descriptions of diagrammatic content. The resulting outputs were then serialized in RDF, enabling machine-readable representations of visual reasoning within complex multimodal sources.

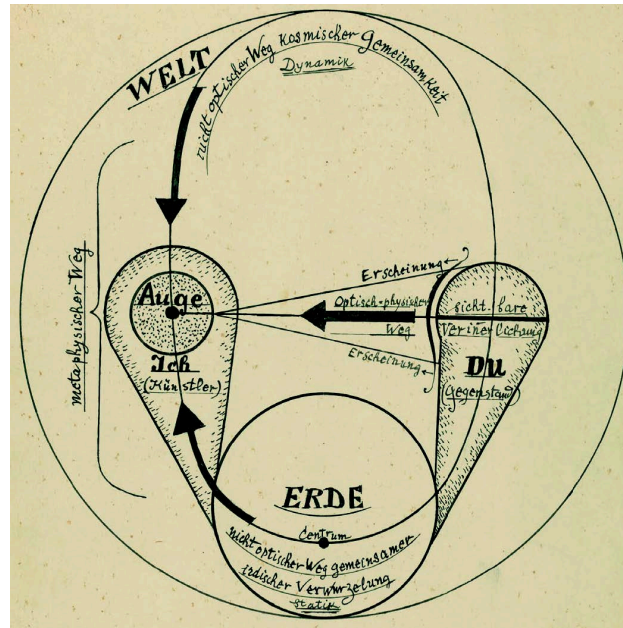


Figure 1: Paul Klee’s Theory of Pictorial Configuration as a diagram. Zentrum Paul Klee, Bern, Inv.Nr. BG A/030. Photo: Zentrum Paul Klee.

2 Background and Related Works

In his work, Peirce often emphasized the role of diagrams in reasoning [32], providing various examples of what he termed *diagrammatic reasoning* (CP 4.571, 5148, 6.213). On the one hand, diagrams exhibit an iconic character: they are a specific subtype of icon capable of representing the internal structure of the objects they depict through the arrangement of their interconnected parts. For example, a map that shows historical trade routes by positioning ports and drawing connecting lines already functions as a diagram, as it represents spatial and relational structures that mirror real-world networks of movement and exchange. On the other hand, diagrams also possess a dynamic character, as they enable manipulation and iterative transformations according to the general laws governing the relationships among their parts, and as such, pose epistemological questions about the generation and production of knowledge [19; 31]. For instance, a simple triangle drawn on paper can serve as a diagram of all triangles by altering its angles or side lengths while preserving its topology. Most notably, Peirce developed a system of visual logic known as *Existential Graphs*, structured into four levels: *Alpha* for propositional logic, *Beta* for modal and higher-order logic, and *Gamma* and *Delta* for meta-assertions and non-declarative statements. *Existential Graphs* use visual connectors such as lines, curves, and nodes to represent logical relations. To start, the *Sheet of Assertion* is a blank space on which diagrams are drawn. Any proposition written directly on it is considered to be asserted as true. A continuous line connecting the elements indicates a



Figure 2: An example of diagram in religious art taken from [12]: Alton Towers Triptych, Cologne (?), ca. 1150. London, Victoria & Albert Museum, inv. no. 4757-1858. Photo: © Victoria & Albert Museum.

logical identity. Enclosures, such as closed curves, represent negation; therefore, a region inside a curve is logically negated. Juxtaposed elements without connectors are logically conjunctive (i.e., both are assumed to be true). A bifurcation indicates a logical disjunction. These characteristics make Peirce’s diagrams interesting for computational modeling; however, their formal and visual complexity also requires methods capable of isolating and structuring heterogeneous visual content within manuscripts.

The structural heterogeneity of historical manuscripts poses significant challenges for automated information extraction and semantic indexing. Recent studies have developed machine learning pipelines for layout segmentation, targeting both textual and non-textual elements, such as illuminations and decorative initials [1; 3]. Object detection models, such as YOLO [33], have also been employed to identify image regions in complex manuscript layouts, offering effective layout segmentation [26]. Among these approaches, Fleischhacker et al. [9] proposed a pipeline that combines layout detection, synthetic data augmentation, OCR fine-tuning, and reintegration of outputs as IIIF-compliant annotations. While such methods enable the scalable processing of visually rich documents, they do not incorporate mechanisms for semantic enrichment or integration into LOD frameworks. To address this issue, ontological extensions of the Web Annotation Data Model (WADM) [30] have been proposed. The Multi-Level Annotation Ontology (MLAO) [23] introduces conceptual anchoring and provenance for annotations, while the General Ekphrastic Ontology (GEkO) [2] models ekphrastic relations between textual descriptions and visual elements. These limitations also highlight the need for approaches that combine visual segmentation with semantic interpretation. In this context, VLMs represent promising tools for bridging the gap between image analysis and knowledge extraction processes.

In recent years, the intersection of computer vision and art history has seen a growing integration of visual and textual modalities, driven by the availability of large art collections of dig-

itized images and the development of multimodal machine learning techniques. This has led to substantial interest in tasks such as visual link retrieval, multimodal classification, iconographic captioning [4], and visual question answering (VQA) [10], particularly in domains such as cultural heritage, where image content is often accompanied by curatorial or scholarly metadata. Although multimodal models can describe visual elements, they often struggle to understand the logical or spatial relationships among them. This limitation has been highlighted in recent studies that show how VLMs tend to rely on background knowledge rather than analyzing the internal structure of diagrams [15]. To overcome this, some approaches combine image segmentation and structured prompting. For example, the chain-of-regions method decomposes diagrams into meaningful areas before interpreting them, improving the model’s reasoning about spatial relations [34]. Other studies have applied VLMs to structured visual domains, such as UML diagrams or flowcharts, using modular reasoning pipelines to achieve more accurate results [21]. However, to the best of the authors’ knowledge, no existing study has specifically addressed the use of VLMs for identifying or interpreting diagrammatic content in historical sources. In this context, applying similar strategies to Peirce’s existential graphs means using layout segmentation to isolate diagram regions and then prompting VLMs with questions informed by Peirce’s semiotics. This structured workflow helps models provide more accurate interpretations of diagrammatic reasoning.

3 Methods

3.1 Corpus Description and Preprocessing

The Charles S. Peirce Papers (MS Am 1632) [28], housed at Harvard’s Houghton Library, represent one of the most extensive archival collections of Peirce’s works. Comprising over 1,700 manuscript items, the collection spans disciplines ranging from mathematics to logic and metaphysics. A subset of 233 items was digitized and made available through IIF Manifests via the Harvard Hollis system [13], yielding a total of 15,695 high-resolution facsimile images.

To prepare the corpus for computational processing, we retrieved IIF metadata for each digitized item, including canvas structure, image URIs, and classification labels derived from Robin’s catalogue [28]. All canvases were downloaded at full resolution and organized into thematic folders. Blank pages, identified using IIF metadata, were automatically excluded, resulting in a set of 13,234 manuscript pages (Table 1).

To contextualize the corpus thematically, we constructed a bump chart showing the distribution of digitized pages in Robin’s topical categories grouped by five-year intervals within Peirce’s lifetime (Figure 3). Category D (Logic) dominates the corpus, followed by Pragmatism (B) and Metaphysics (E), reflecting Peirce’s focus on formal reasoning and motivating our attention to visual content in these areas.

Description	Count
Total manuscript items	1,759
Digitized items	233
Total digitized pages	15,695
Blank pages removed	2,461
Pages retained for processing	13,234

Table 1: Key statistics of the Peirce manuscript corpus.

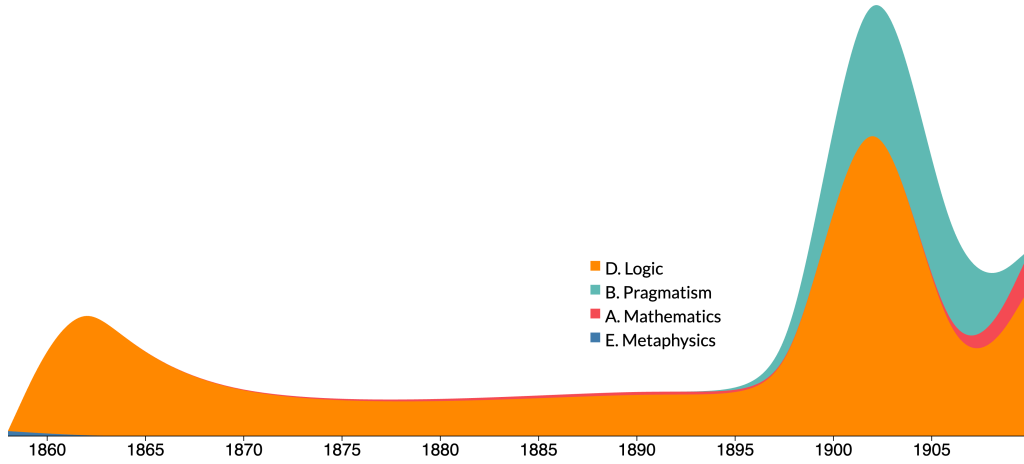


Figure 3: Distribution of digitized manuscript pages across Peirce’s lifetime, grouped by five-year intervals and categorized according to Robin’s classification. The visualization, based on IIIF canvas data, highlights how Peirce’s intellectual focus evolved over time, with Logic manuscripts dominating the corpus and peaks corresponding to his most productive years in formal reasoning.

3.2 Models for Page and Layout Analysis

To distinguish textual from visually mixed pages, we implemented a classification pipeline based on three feature extraction strategies: Histogram of Oriented Gradients (HOG), intermediate features from ResNet18, and semantic embeddings from the CLIP visual encoder. Each page was labeled as either *text* (pages containing mostly textual elements), *diagram_mixed* (pages containing at least one relevant visual feature), or *cover* using a manually annotated dataset of 1,264 pages.

To extract diagrammatic content from the classified *diagram_mixed* pages, we created a manually annotated dataset of 443 manuscript images. Each page was labeled using a two-class schema: *diagram* and *text_block*. This typology was designed to identify the main visual and textual regions while preserving the layout-level structure and supporting generalization across diverse manuscript formats. At this stage, we intentionally avoided adding more granular annotations, such as dates, titles, sketches, logical notation, or algebraic formulas, to prevent class imbalances in the training data.

The dataset was split into training and validation subsets using an 80/20 ratio. To address the class imbalance and increase robustness, we applied data augmentation to all pages containing at least one diagram annotation. Two synthetic variants were generated for each page, resulting in a final training set of 1,133 images. Finally, we fine-tuned the YOLOv8m model on this dataset, which was selected for its balance between detection performance and computational efficiency.

3.3 Annotation Workflow

All detected segments, whether diagrams or text blocks, were transformed into structured annotations compliant with the IIIF. Each segment is expressed as an instance of WADM, linking a specific region of the manuscript image (the *target*) to a content resource or metadata element (the *body*), serialized in JSON-LD format. Bounding boxes are mapped to IIIF Canvas coordinates using the *xywh* fragment selector, ensuring the precise anchoring of visual elements within the page layout.

To enhance semantic expressiveness, we employ MLA0 [23], an extension of WADM. The MLA0 introduces the *mlao:Anchor* class to separate the annotated physical region from its conceptual referent. In our use case, anchors are linked to the IIIF URI of the full manuscript page via *mlao:isAnchoredTo*, enabling a shared conceptual reference for both textual and diagrammatic

segments of the manuscript. Instead of predefined abstraction layers (e.g., Work, Expression, Manifestation, and Item according to LRMoo [27]), we define custom conceptual categories based on Peirce’s semiotic theory (see §3.4). Interpretative captions generated by the VLM are modeled as `oa:TextualBody` instances linked to a `hico:InterpretationAct` [8] that specifies the interpretative level, model used, and generation process via PROV-O. This structure supports the hermeneutic traceability and versioning of automated interpretations.

Annotations are generated from the detection outputs and can be embedded in IIIF Manifests or published as standalone pages. Annotations can also be serialized in RDF for semantic querying. This semantic layer supports integration into LOD workflows and prepares the content for VLM interpretation and use.

3.4 VLM Prompting and Interpretation

Peirce’s semiotic theory offers a framework to understand the structure and function of signs, which we use as a basis to design VLMs prompts. We define three analytical categories for prompt engineering that operationalize aspects of Peirce’s semiotic theory for computational analysis. These categories are designed for VLM prompting rather than direct applications of his icon-index-symbol trichotomy. The morphological level addresses the basic visual elements that constitute a diagram, such as lines, shapes, and symbols. This corresponds broadly to the iconic mode of representation and relates to Peirce’s category of Firstness. The indexical level concerns the relationships between these elements, identifying connections, dependencies, or structural links. This reflects aspects of Peirce’s notion of Secondness. The symbolic level explores the logical operations encoded in the diagram. At this stage, we provide the VLM with minimal instructions on how to interpret visual conventions (e.g., enclosure, juxtaposition, lines of identity), prompting the model to reconstruct the inferential logic underlying the structure. This aligns with Peirce’s category of Thirdness. Table 2 shows the specific question templates for each category.

Finally, we defined three classes extending the MLA0 data model (`pip:MorphologicalLevel`, `pip:IndexicalLevel`, and `pip:SymbolicLevel`) to represent the semiotic categories described earlier. The VLM-generated responses, along with metadata about the model and prompt, were re-injected as annotations into the IIIF-compliant JSON-LD structure. Each annotation targets a specific region of the IIIF canvas and references the full manuscript page via its persistent URI.

Semiotic Level	Question Template
Morphological	How many and what kind of elements (e.g., words, lines, arcs, nodes, shapes, etc.) are present in the image?
Indexical	Is there a relationship between the elements present in the image? Which elements are connected to each other?
Symbolic	<p>In Peirce’s diagrammatic logic, a closed curve called a <i>cut</i> represents logical negation. Elements inside the same region are interpreted conjunctively (i.e., asserted together). Elements placed directly on the background (the Sheet of Assertion) are considered true. A cut around propositions denies them. Nested cuts represent nested negation. Lines may indicate identity or existential quantification.</p> <p>Based on these principles, interpret the diagram and translate its meaning into a logical statement. If this is not possible, provide a clear explanation in natural language.</p>

Table 2: Template of VLM Questions Based on Semiotic Categories

3.5 Evaluation Methodology

To assess the interpretative capabilities of VLMs with respect to Peirce’s diagrammatic logic, we conducted a qualitative evaluation across five diagrams of increasing complexity, manually selected from the Peirce manuscript corpus and belonging to the Alpha level. Standard reference-based metrics, such as CLIPScore [14] are limited in this context, as they primarily measure lexical similarity and do not account for the semantic or structural accuracy of a caption [5; 7]. This makes them unsuitable for evaluating the descriptions of abstract and diagrammatic content. For each diagram, three structured prompts were submitted to the models, corresponding to Peirce’s semiotic categories: morphological (element enumeration), indexical (relational structure), and symbolic (logical translation), as shown in Table 2. We tested five VLMs: **BLIP3-o** [6], **GPT-4o**, **LLaVA 1.6 vicuna-13b** [22], **MiniGPT-4 vicuna-13b** [35] and **Phi-4 Multimodal** [11], all of which accept both visual and textual inputs. This evaluation aimed to compare their capacity to recover structured meaning from diagrammatic forms, with particular attention to inferential depth and semiotic coherence.

Each response was rated on a 3-point scale: 2 for correct and complete answers, 1 for partially correct answers, and 0 for incorrect or irrelevant responses. The total possible score was 30 (3 questions \times 5 diagrams \times 2 points).

4 Results and Discussion

4.1 Performance of Preparatory Models

The best performance for image classification was achieved using a logistic regression classifier trained on CLIP embeddings. Using 10-fold stratified cross-validation, the model achieved a macro-averaged F1-score of **0.9531**, with good class-wise accuracy across the board. The full results of the model comparison are reported in Appendix A. To assess the distribution of visual content across the corpus, we applied the trained classifier to all digitized pages and aggregated the predictions according to thematic categories. As shown in Figure 4, visual content is especially concentrated in Category D (Logic), followed by Pragmatism (B) and Metaphysics (E), suggesting that Peirce used visual reasoning more frequently in these areas.

Class	Precision	Recall	F1 Score	Support
Cover (0)	1.0000	1.0000	1.0000	28
Text (1)	0.9459	0.8974	0.9211	117
Diagram (2)	0.9040	0.9496	0.9262	119

Table 3: Performance of the best model (Logistic Regression + CLIP) by class.

On the validation set (111 pages), the fine-tuned YOLOv8m model achieved a mean Average Precision at IoU 0.5 (mAP@0.5) of **0.981**, with a class-specific score of **0.992** for diagram and **0.970** for text_block. The precision, recall, and F1 scores are reported in Table 4. A sample prediction with annotations is shown in Figure 5.¹

4.2 Preliminary VLM Evaluation

As shown in Table 5, GPT-4o achieved the highest score (25/30), demonstrating relatively strong performance across all semiotic dimensions. BLIP3-o followed with 12/30, showing partial competence in recognizing visual elements but struggling with relational and symbolic interpretation.

¹ The models and the scripts used for the preprocessing and evaluation are available at: <https://anonymous.4open.science/r/PIP-Manuscripts-Processor-0147/>.

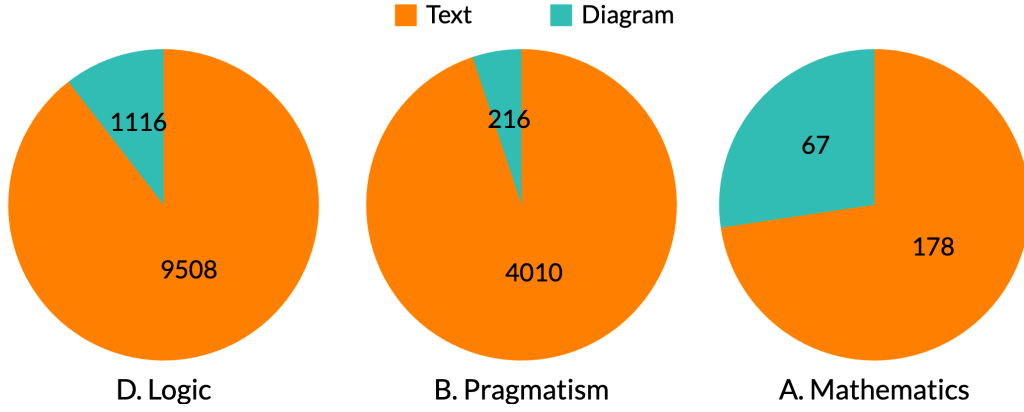


Figure 4: Pie charts show text and diagrams distribution within three categories.

Metric	Diagram	Text Block	All Classes
mAP@0.5	0.992	0.970	0.981
Precision (at best F1)	0.990	0.960	0.975
Recall (at best F1)	1.000	0.940	0.990
Optimal F1 score	0.996	0.948	0.960
Confidence @ Optimal F1	0.547	0.547	0.547
Confidence @ Max Precision	0.975	0.975	0.975
Confusion (TP)	684	473	—
Confusion (FP)	38	58	—
Confusion (FN)	3	23	—

Table 4: Detection performance on the Peirce manuscript validation set (n=111 images).

Phi-4 obtained a total of 6 points, with modest success in symbolic recognition but weak results in the other dimensions. Both LLaVA 1.6 and MiniGPT-4 scored 0, failing to produce meaningful responses to any of the evaluative questions.

While GPT-4o and BLIP3-o can handle layout-level tasks without fine-tuning, their performance drops significantly when confronted with more complex diagrams involving nested cuts or non-trivial spatial configurations. Some errors are attributable to OCR-like misrecognition, such as reading “wounded” as “mound” or “man” as “noun”. Across nearly all models, the symbolic level obtained the lowest average scores, with frequent failures in understanding negation correctly (e.g., misinterpreting a cut as emphasis, ignoring it entirely) and generating logically valid formalizations (e.g., confusing $\neg(A \wedge B)$ with $\neg A \wedge \neg B$), or even hallucinating logical rules (in smaller models). Interestingly, considering the diagram in Figure 6, both GPT-4o and BLIP3-o produced formal logical statements in response to the symbolic question. GPT-4o correctly generated: “There exists a man who is not both wounded and disgraced,” and formalized it as:

$$\exists x (\text{Man}(x) \wedge \neg (\text{Wounded}(x) \wedge \text{Disgraced}(x))) . \quad (1)$$

BLIP3-o instead produced: “It is not the case that there exists a man who is wounded and disgraced,” rendered as:

$$\neg \exists x (\text{Man}(x) \wedge \text{Wounded}(x) \wedge \text{Disgraced}(x)) . \quad (2)$$

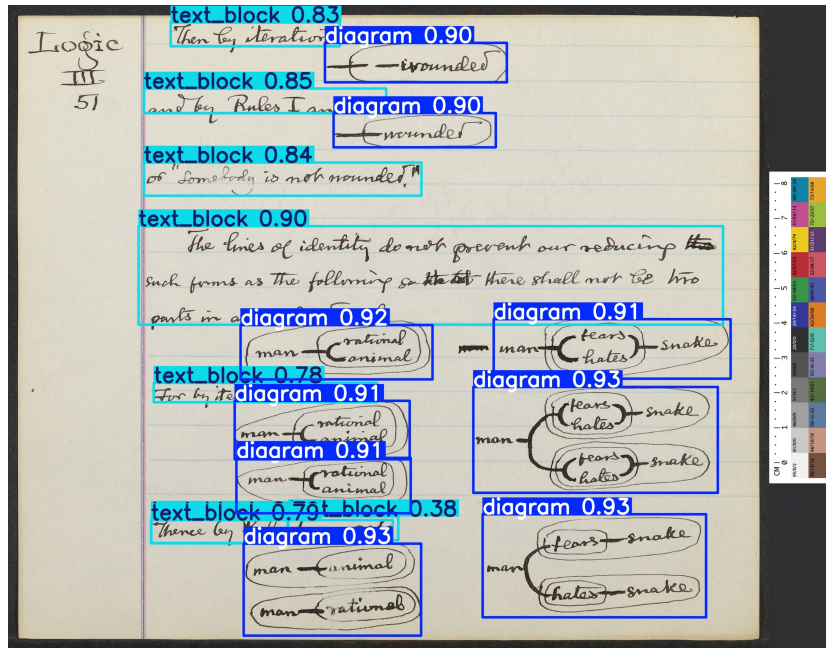


Figure 5: Output of the fine-tuned YOLOv8m model on autograph manuscript dated 1902. MS Am 1632 (430), Box 29, Folder 28, Series I. Manuscripts, D. Logic. Houghton Library, Harvard University, USA. Photo: © Houghton Library. Persistent URL: <http://nrs.harvard.edu/urn-3:FHCL.HOUGH:12491033>. The model correctly identifies and segments 'diagram' (blue) and 'text_block' (light blue) regions, providing the structured data used for subsequent annotation and VLM analysis.

This suggests that while BLIP3-o demonstrates surface-level competence in formalization, GPT-4o is better able to align visual features with underlying logical relations.

To finalize the workflow suggested at the end of §3.4, we also produced the RDF serialization of the annotations generated by the model.²

Model	Morphological	Indexical	Symbolic	Total
GPT-4o	7	9	9	25
BLIP3-o	3	5	4	12
Phi-4	1	1	4	6
LLaVA 1.6	0	0	0	0
MiniGPT-4	0	0	0	0

Table 5: Qualitative evaluation scores across five diagrams. Maximum: 30 points per model.

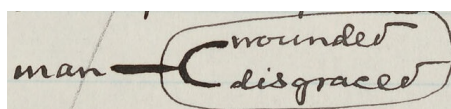


Figure 6: Example diagram from Peirce's *Logic* manuscripts, used for evaluating the interpretative capacity of VLMs.

² The dataset and RDF serializations are available at <https://doi.org/10.5281/zenodo.16113285>.

The high concentration of diagrams in Peirce’s manuscripts offers a quantitative overview of the importance of visual representations as working instruments during reasoning itself, aligning with the philosopher pragmatic maxim that concepts acquire meaning through their operational use rather than from correspondence to abstract or *a priori* definitions. In other words, Peirce resorted to diagrammatic reasoning on paper as a privileged way to develop his arguments and to enable the manipulation of relational structures. Through this perspective, this exploratory analysis reveals Peirce’s philosophy as an embodied practice where “moving pictures of thought” and linear writing operate in a continuous exchange, underlining at the same time the importance of intermediality in such heterogeneous manuscripts, supporting the arguments raised by Keeler [18]. Our preliminary VLM evaluation serves as a pilot study. Expanding the dataset would enable systematic investigation of how Peirce’s model of semiosis relates to how VLMs process diagrammatic content, and whether studying these models can in turn clarify what diagrammatic reasoning requires. The workflow presented here makes such evaluation feasible across the full corpus.

5 Conclusion

This preliminary work presents a modular pipeline for analyzing heterogeneous manuscript collections, combining layout classification, object detection, and semantic annotation within IIIF and LOD frameworks. The workflow extends WADM through MLAO and introduces a qualitative VLM evaluation method structured around analytical categories derived from Peirce’s semiotic theory. Applied to Peirce’s manuscripts, the analysis reveals the quantitative distribution shows that visual reasoning is concentrated in specific philosophical domains, with Logic manuscripts containing 10.5% diagrams versus 5.1% in Pragmatism. This shows that diagrammatic practice was functionally integrated into Peirce’s work on formal systems. Second, VLM evaluation reveals that GPT-4o can approximate logical interpretation when appropriately prompted (25/30 points), while smaller models underperform. The differential performance across morphological, indexical, and symbolic questions validates these as functionally distinct analytical operations.

The methodological pattern extends beyond Peirce studies. Any manuscript collection combining text and visual elements can adapt this workflow by substituting domain-appropriate theoretical frameworks, retraining segmentation models, and adjusting prompts to specific research questions. Moreover, the workflow can be integrated into semantic digital editions based on digital facsimiles. Textual content can be transcribed using HTR, while annotations contextualize visual elements at multiple scales. The annotations generated through this process can automatically enrich knowledge graphs. Finally, by treating annotations as digital traces of interpretation, the system aligns with Peirce’s pragmatist view of semiosis as an ongoing process.

References

- [1] Aouinti, Mehdi, Plancq, Clément, Froeliger, Nicolas, Leclère, François, and Rico, Christophe. “Detecting Illuminations in Digitized Medieval Manuscripts: A Benchmark Dataset and Evaluation Protocol”. In: *Digital Medievalist* 15, no. 1 (2022), pp. 1–18. DOI: 10.16995/dm.7844.
- [2] Bocchi, Maria Francesca, Pedretti, Carlo Teo, and Vitali, Fabio. “Between Text and Icon: Towards A Representational Model for Ekphrastic Relations”. In: *Proceedings del XIV Convegno Annuale AIUCD2025*, ed. by Simone Rebora, Marco Rospocher, and Stefano Bazaco. Verona, Italy: AIUCD, 2025, pp. 566–572. DOI: 10.6092/UNIBO/AMSACTA/8380.
- [3] Büttner, Stefan, Wettig, Tilo, König, Dominik, Springer, Matthias, and Wittler, Roland. “CorDeep: A Deep Learning-Based Approach to the Detection and Classification of Visual Elements in Historical Documents”. In: *Journal of Imaging* 8, no. 10 (2022), pp. 285–303. DOI: 10.3390/jimaging8100285.

- [4] Cetinic, Eva. "Towards Generating and Evaluating Iconographic Image Captions of Art-works". In: *Journal of Imaging* 7, no. 8 (2021), p. 123. DOI: 10.3390/jimaging7080123.
- [5] Cetinic, Eva, Lipic, Tomislav, and Grgic, Sonja. "Fine-tuning convolutional neural networks for fine art classification". In: *Expert Systems with Applications* 114 (2018), pp. 107–118.
- [6] Chen, Jiuhai, Xu, Zhiyang, Pan, Xichen, Hu, Yushi, Qin, Can, Goldstein, Tom, Huang, Lifu, et al. "BLIP3-o: A Family of Fully Open Unified Multimodal Models–Architecture, Training and Dataset". arXiv preprint arXiv:2505.09568. 2025. DOI: 10.48550/ARXIV.2505.09568. arXiv: 2505.09568 [cs.CL].
- [7] D’Armenio, Enzo, Deliège, Adrien, and Dondero, Maria Giulia. "A Semiotic Methodology for Assessing the Compositional Effectiveness of Generative Text-to-Image Models (Mid-journey and DALL·E)". In: *Lecture Notes in Computer Science*. Cham: Springer Nature Switzerland, 2025, pp. 112–127. DOI: 10.1007/978-3-031-92089-9_8.
- [8] Daquino, Marilena and Tomasi, Francesca. "Historical Context Ontology (HiCO): A Conceptual Model for Describing Context Information of Cultural Heritage Objects". In: *Metadata and Semantics Research*, ed. by Emmanouel Garoufallou, Richard J. Hartley, and Panorea Gaitanou. Cham: Springer International Publishing, 2015, pp. 424–436. DOI: 10.1007/978-3-319-24129-6_37.
- [9] Fleischhacker, David, Göderle, Wolfgang Thomas, and Kern, Roman. "Text Extraction for Complex Historical Documents: A Modular Approach to Layout Detection and OCR". In: *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery, 2025, pp. 1–3. DOI: 10.1145/3677389.3702524.
- [10] Garcia, Noa and Vogiatzis, George. "How to read paintings: Semantic art understanding with multi-modal retrieval". In: *Computer Vision – ECCV 2018 Workshops, Proceedings*, ed. by Stefan Roth and Laura Leal-Taixé. Vol. 11130. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, 2019, pp. 676–691. DOI: 10.1007/978-3-030-11012-3_52.
- [11] Haider, Emman, Perez-Becker, Daniel, Portet, Thomas, Madan, Piyush, Garg, Amit, Ashfaq, Atabak, Majercak, David, et al. "Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle". arXiv preprint arXiv:2407.13833. 2024. DOI: <https://doi.org/10.48550/arXiv.2407.13833>. arXiv: 2407.13833 [cs.CL].
- [12] Hamburger, Jeffrey F. *Diagramming Devotion: Berthold of Nuremberg’s Transformation of Hrabanus Maurus’s Poems in Praise of the Cross*. Chicago: University of Chicago Press, 2019. DOI: 10.7208/chicago/9780226642956.001.0001.
- [13] Harvard University. "Charles S. Peirce Papers". <https://hollisarchives.lib.harvard.edu/repositories/24/resources/6437>. Accessed: 2025-04-04. 2023.
- [14] Hessel, Jack, Holtzman, Ari, Forbes, Maxwell, Le Bras, Ronan, and Choi, Yejin. "CLIP-Score: A Reference-Free Evaluation Metric for Image Captioning". arXiv preprint arXiv:2104.08718. 2021. DOI: 10.48550/ARXIV.2104.08718. arXiv: 2104.08718 [cs.CV].
- [15] Hou, Yifan, Giledereli, Buse, Tu, Yilei, and Sachan, Mrinmaya. "Do Vision-Language Models Really Understand Visual Language?" In: *arXiv* (2024). DOI: 10.48550/ARXIV.2410.00193. URL: <https://arxiv.org/abs/2410.00193>.
- [16] Keeler, Mary. "Iconic Indeterminacy and Human Creativity in the C.S. Peirce’s Manuscripts". In: *The Iconic Page in Manuscript and Digital Culture*, ed. by George Bornstein and Theresa L. Tinkle. Ann Arbor, MI: University of Michigan Press, 1998, pp. 157–194.

- [17] Keeler, Mary. “Pragmatically Improving Access to Peirce’s Archive”. In: *Chinese Semiotic Studies* 16, no. 1 (2020), pp. 167–187. DOI: 10.1515/css-2020-0009.
- [18] Keeler, Mary. “The Hidden Treasure of C. S. Peirce’s Manuscripts”. In: *Chinese Semiotic Studies* 16, no. 1 (2020), pp. 155–166. DOI: 10.1515/css-2020-0008.
- [19] Kiryushchenko, Vitaly. *Diagrams, Visual Imagination, and Continuity in Peirce’s Philosophy of Mathematics*. 1st ed. Mathematics in Mind Series. Cham: Springer International Publishing AG, 2023.
- [20] Latour, Bruno. “Visualisation and Cognition: Drawing Things Together”. In: *Knowledge and Society: Studies in the Sociology of Culture Past and Present*, ed. by Henrika Kuklick. Vol. 6. JAI Press, 1990, pp. 1–40.
- [21] Liang, Yichi et al. “FlowLearn: A Modular Framework for Diagram Understanding”. arXiv preprint arXiv:2412.16420v1. 2024. arXiv: 2412.16420v1.
- [22] Liu, Haotian, Li, Chunyuan, Wu, Qingyang, and Lee, Yong Jae. “Visual Instruction Tuning”. arXiv preprint arXiv:2304.08485. 2023. DOI: 10.48550/arXiv.2304.08485. arXiv: 2304.08485 [cs.CV].
- [23] Pedretti, Carlo Teo, Bocchi, Maria Francesca, Tomasi, Francesca, and Vitali, Fabio. “What Do We Annotate When We Annotate? Towards A Multi-Level Approach to Semantic Annotations”. In: *Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI*, ed. by Angelo A. Salatino, Mehwish Alam, Femke Ongenaes, Sahar Vahdati, Anna Lisa Gentile, Tassilo Pellegrini, and S. Jiang. Vol. 60. Studies on the Semantic Web. Amsterdam: IOS Press, 2024, pp. 370–385. DOI: 10.3233/SSW240030.
- [24] Peirce, Charles S. *Collected Papers of Charles Sanders Peirce, Vols. 1-6*, ed. by Charles Hartshorne and Paul Weiss. Cambridge, MA: Harvard University Press, 1931.
- [25] Picca, Davide, Schnyder, Aurélie, Kostina, Ekaterina, Adamou, Anastasia, Rodighiero, Dario, and Schnapp, Jeffrey. “Orchestrating Cultural Heritage: Exploring the Automated Analysis and Organization of Charles S. Peirce’s PAP Manuscript”. In: *Proceedings of the 34th ACM Conference on Hypertext and Social Media (HT ’23)*. Rome, Italy: Association for Computing Machinery, 2023, pp. 1–4. DOI: 10.1145/3603163.3609066.
- [26] Ravichandra, S., Sathya, S. Siva, and Sophie, L. “Deep Learning Based Document Layout Analysis on Historical Documents”. In: *Advances in Distributed Computing and Machine Learning*. Singapore: Springer, 2022, pp. 73–85. DOI: 10.1007/978-981-16-9465-1_8.
- [27] Riva, P., Žumer, M., and Aalberg, T. “LRMoo, a High-Level Model in an Object-Oriented Framework”. Report. Available at <https://repository.ifla.org/handle/20.500.14598/2217>. IFLA, 2022.
- [28] Robin, Richard S. “The Peirce Papers: A Supplementary Catalogue”. In: *Transactions of the Charles S. Peirce Society* 7, no. 1 (1971), pp. 37–57.
- [29] Rodighiero, Dario, Romele, Alberto, Higuera Rubio, José, Pedro, Celeste, Azzi, Matteo, and Uboldi, Giorgio. “Advanced Interface Design for IIIF. A Digital Tool to Explore Image Collections at Different Scales”. In: *Umanistica Digitale*, no. 15 (2023), pp. 167–192. DOI: 10.6092/ISSN.2532-8816/17230.
- [30] Sanderson, Robert, Ciccicarese, Paolo, and Young, Benjamin. “Web Annotation Data Model”. W3C Recommendation. <https://www.w3.org/TR/annotation-model/>. W3C, 2017.

- [31] Frederik Stjernfelt, edited by. *Diagrammatology: An Investigation On The Borderlines Of Phenomenology, Ontology, And Semiotics*. Red. by Vincent F. Hendricks, John Symons, Jaakko Hintikka, Dirk Van Dalen, Theo A.F. Kuipers, Teddy Seidenfeld, Patrick Suppes, and Jan Woleński. Vol. 336. Synthese Library. Dordrecht: Springer Netherlands, 2007. DOI: 10.1007/978-1-4020-5652-9. URL: <https://link.springer.com/10.1007/978-1-4020-5652-9> (visited on 10/25/2025).
- [32] Waal, Cornelis de. *Peirce: A Guide for the Perplexed*. Guides for the Perplexed. London: Bloomsbury Publishing, 2013. ISBN: 9781847065155.
- [33] Yaseen, Muhammad. "What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector". 2024. DOI: 10.48550/arXiv.2408.15857. arXiv: 2408.15857 [cs.CV]. URL: <https://arxiv.org/abs/2408.15857>.
- [34] Zhang, Ruohong, Zhang, Bowen, Li, Yanghao, Zhang, Haotian, Sun, Zhiqing, Gan, Zhe, Yang, Yinfei, Pang, Ruoming, and Yang, Yiming. "Improve Vision Language Model Chain-of-thought Reasoning". 2024. arXiv: 2410.16198 [cs.AI]. URL: <https://arxiv.org/abs/2410.16198>.
- [35] Zhu, Deyao, Chen, Jun, Shen, Xiaoqian, Li, Xiang, and Elhoseiny, Mohamed. "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models". 2023. DOI: 10.48550/arXiv.2304.10592. arXiv: 2304.10592 [cs.CV]. URL: <https://arxiv.org/abs/2304.10592>.

A Model Comparison for Page Classification

Feature	Model	Avg Precision	Avg Recall	Avg F1 Score	Accuracy
CLIP	Linear SVM	0.9335	0.9320	0.9321	0.9091
CLIP	Logistic Regression	0.9500	0.9490	0.9491	0.9318
CLIP	Random Forest	0.9296	0.9293	0.9294	0.9053
CLIP	SVM RBF	0.9500	0.9460	0.9461	0.9280
CLIP	k-NN (k=5)	0.9123	0.9093	0.9093	0.8788
CNN	Linear SVM	0.9351	0.9350	0.9350	0.9129
CNN	Logistic Regression	0.9356	0.9349	0.9350	0.9129
CNN	Random Forest	0.9183	0.9180	0.9181	0.8902
CNN	SVM RBF	0.9183	0.9180	0.9181	0.8902
CNN	k-NN (k=5)	0.8932	0.8835	0.8882	0.8561
HOG	Linear SVM	0.8341	0.8243	0.8290	0.7765
HOG	Logistic Regression	0.8361	0.8361	0.8361	0.7803
HOG	Random Forest	0.8688	0.8375	0.8504	0.8182
HOG	SVM RBF	0.8548	0.8351	0.8442	0.8030
HOG	k-NN (k=5)	0.6659	0.6974	0.5889	0.6212

Table 6: Performance summary of all models across features. Best values are in bold.

True / Predicted	Cover	Text	Diagram_mixed
Cover	28	0	0
Text	0	105	12
Diagram_mixed	0	6	113

Table 7: Confusion matrix for Logistic Regression + CLIP embeddings (aggregated over 10 folds).