DeepVis: Directed Multimodal Preprocessing for Document Understanding Beyond Text Extraction

Christopher Robert Price

***Abstract*—Research across visually-rich document collections demands multimodal understanding that text-centric retrieval systems cannot provide. Retrieval-Augmented Generation (RAG) relies on OCR-based text extraction, discarding visual and spatial information critical for disciplines such as art history, while semantic chunking produces non-deterministic, fragmented context that obscures page-level structure and increases hallucination risk. We present DeepVis, a two-stage multimedia processing pipeline addressing these limitations through vision-based preprocessing, deterministic context organization, and exact page citation. Vision-capable LLMs preprocess documents page-by-page into directed artifacts organized in strict document order, preserving textual, visual, and spatial information as coherent, reproducible memory.**

We evaluate DeepVis against Azure Cognitive Search RAG on Owen Jones's *The Grammar of Ornament*, a 612-page multimodal corpus of hand-lithographed ornamental plates, using 15 queries across three stratified categories. DeepVis achieved 100% Page Citation Recall versus RAG's 6.7% and 80% Visual Content Accuracy versus RAG's 0%; nominally correct RAG visual responses were derived from co-located text, confirming the absence of visual reasoning capability. Query success rate was 67% versus 40%, with failures attributable to static preprocessing. Deployed university-wide following pilot testing, DeepVis has expanded to administrative and public-facing applications, demonstrating that directed multimodal preprocessing enables document understanding that text-centric retrieval architectures fundamentally cannot support.

***Index Terms*—Multimodal Document Understanding, Vision-Language Models, Multimedia Information Retrieval, Document Image Analysis, Directed Multimodal Comprehension.**

I. Introduction

MULTIMODAL document understanding presents a fundamental challenge for automated research systems: large collections of visually-rich materials carry meaning across text, image, and spatial arrangement simultaneously, yet the dominant retrieval architectures treat documents as text alone. In disciplines where visual content is primary evidence, such as art history, the inability to reason across modalities is not a performance limitation but a categorical one. A system that cannot analyze an image cannot answer a question whose answer lives in that image, regardless of how sophisticated its text retrieval is.

Retrieval-Augmented Generation (RAG) has become the dominant paradigm for grounding large language model (LLM) responses in external knowledge sources [1]. RAG systems create embeddings of document chunks, store them in vector databases, and retrieve semantically relevant fragments at query time to augment LLM responses. While effective for text-based question answering, RAG faces three fundamental limitations when applied to visually-rich documents. First, RAG systems extract only text through OCR, discarding visual and spatial information such as image composition, color relationships, diagrammatic structure, and page layout. Second, semantic similarity search returns chunks in non-deterministic order based on vector distances and index state, meaning the same query can produce different chunk orderings as the index grows or retrieval parameters change, contributing to inconsistent responses and hallucination. Third, the chunking process fragments page-level structure, making accurate page citation unreliable, which presents a critical limitation for research requiring source verification.

These limitations became concrete when researchers at the University of California Irvine sought to analyze thousands of pages of art history material. Queries such as "Find all instances of Surrealism" require visual comprehension of artworks, coherent analysis across large document collections, and precise page references for deeper investigation. Initial attempts with Microsoft Azure Cognitive Search failed to support this use case: the system produced incomplete results without reliable page citations and lacked the visual awareness necessary for art historical analysis, frequently generating hallucinations and "I don't know" responses when visual content was the primary evidence.

We present DeepVis, a two-stage multimedia processing pipeline that addresses these limitations through three contributions to multimodal document understanding. First, *vision-based preprocessing* leverages vision-capable LLMs to analyze each document page as a multimodal input, generating user-directed artifacts that preserve textual, visual, and spatial information according to per-document prompts specified at upload time. Second, *deterministic context organization* arranges these artifacts in strict document and page order within the inference model's context space, creating coherent document memory that produces consistent, reproducible results and limits hallucination risk. Third, *exact page references* maintained throughout preprocessing and inference allow users to verify system responses and conduct deeper investigation with confidence.

We evaluate DeepVis on Owen Jones's *The Grammar of Ornament* [12], a 612-page canonical art history reference comprising 100 hand-lithographed color plates spanning ornamental traditions from ancient Egyptian through Renaissance European design. Using 15 queries stratified across text-accessible, visually-dependent, and cross-modal synthesis

Christopher Robert Price is with the Office of Information Technology, University of California at Irvine, Irvine, CA 92697 USA (e-mail: crprice@uci.edu).

Manuscript received March 2026.

categories, we demonstrate that DeepVis achieves 100% Page Citation Recall compared to RAG's 6.7%, and 80% Visual Content Accuracy compared to RAG's 0%. Analysis of RAG's nominally correct visual responses reveals they were derived from co-located text rather than image content, confirming that OCR-based preprocessing provides no path to visual reasoning regardless of query success rate. DeepVis's Category M underperformance, traced to general-purpose rather than domain-specific preprocessing, identifies the primary current limitation and motivates the active development roadmap.

Developed in August 2024 using GPT-4o's vision capabilities and later migrated to GPT-4.1, DeepVis was piloted in an art history research seminar from September through November 2024. Following critical and constructive feedback, the system was refined and deployed through ClassChat, UCI's ZotGPT classroom tool built on DeepVis preprocessing [13]. It later expanded from academic research to administrative assistants and public-facing chatbots. This deployment trajectory validates the generality of directed multimodal preprocessing beyond the academic use case that motivated its development.

The main contributions of this paper are:

- A two-stage multimodal document understanding pipeline in which user-specified prompts guide vision LLM preprocessing page-by-page, preserving textual, visual, and spatial information without task-specific training or fine-tuning, with one-time analysis cost enabling unlimited cached query reuse.
- Deterministic context organization that arranges preprocessed artifacts in strict document and page order, limiting hallucination risk and enabling exact page citation unavailable to chunk-based retrieval systems.
- Empirical evaluation on a multimodal art history corpus demonstrating 80% Visual Content Accuracy and 100% Page Citation Recall, compared to 0% and 6.7% respectively for Azure Cognitive Search RAG, with analysis of failure modes motivating an active development roadmap.

II. Related Work

A. Retrieval-Augmented Generation Systems

Retrieval-Augmented Generation (RAG) has become the dominant paradigm for grounding large language model responses in external knowledge sources. The standard RAG architecture operates in two phases: an indexing phase that chunks documents, generates embeddings using encoder models, and stores these vectors in specialized databases; and a retrieval phase that embeds user queries, performs semantic similarity search to identify relevant chunks, and augments the LLM prompt with retrieved chunks. This approach has proven effective for text-based question answering over large document corpora, enabling LLMs to access information beyond their training data [1].

However, RAG systems typically rely on OCR-based text extraction, discarding the visual and spatial information critical for understanding documents like art history material, scientific papers with complex diagrams, or near illegible handwriting. Even when OCR successfully extracts text, the semantic chunking and embedding process fragments context across multiple pieces, leading to several fundamental limitations. First, semantic similarity search returns chunks in non-deterministic order based on vector distances, meaning the same query can produce different chunk orderings as the index grows, is replicated for scaling performance, or when retrieval parameters change, contributing to inconsistent responses and hallucination. Recent work has shown that parallel unordered contexts in RAG systems exacerbate both fact fabrication and fact omission [2]. Second, the chunking process alters page-level structure, making it difficult to provide accurate page citations for verification, a critical requirement for academic research. Third, the loss of spatial relationships means systems cannot understand, for example, how a caption relates to its image or how a diagram's layout conveys meaning.

Commercial RAG implementations such as Microsoft Azure Cognitive Search[1] [3] have made RAG more accessible but inherit these core limitations when applied to visually-rich documents, as confirmed in our initial art history deployment.

Recent advances address RAG's limitations through increasingly complex architectures: XSum [4], a modular pipeline for scientific literature summarization, adds specialized question-generation and editor modules that improve retrieval relevance and summary coherence over baseline RAG. Evaluated on the SurveySum dataset, XSum achieves improvements across all evaluated metrics including ROUGE, BERTScore, G-Eval, CheckEval, and Ref-F1. It acknowledges persistent challenges including low abstractive quality scores, verbosity, and cost-effective scaling to large datasets — though scalability to real-world deployment settings is identified as an open direction for future work. This pattern of growing pipeline complexity to address RAG's fundamental limitations motivates DeepVis's alternative strategy of directed preprocessing rather than retrieval refinement.

B. Vision-Capable Large Language Models

Vision-language models enable a qualitatively different approach to document analysis by processing page images holistically [5], enabling simultaneous reasoning over text, spatial layout, and visual content in a single forward pass [6]. Traditional layout analysis systems use sequential rule-based stages [8], while deep learning OCR systems [10] focus on text extraction; even layout-aware pre-training approaches [9] that augment OCR-extracted text with 2D position embeddings and word-level image patches remain constrained by OCR dependency, processing cropped regions rather than whole pages and requiring task-specific fine-tuning. Both paradigms discard compositional arrangement, color semantics, figure-text relationships, and visual pattern geometry carrying primary meaning in multimodal documents. Vision-language models instead comprehend these properties directly, how captions relate to figures, how diagrams encode information, and how visual features contribute to semantic content, with no task-specific training required [5], [6]. Advances in context

[1]Microsoft rebranded *Azure Cognitive Search* to *Azure AI Search* [3]. The cited source uses the current product name *Azure AI Search* throughout. This paper retains *Azure Cognitive Search* to reflect the branding in use at the time of the described deployment.

window capacity and instruction following have extended their practical applicability to multi-page document collections [7].

Two deployment strategies exist for applying vision-language models to document analysis. Real-time processing analyzes page images at query time, adapting analysis to each specific question but incurring per-query inference costs that scale with collection size and query frequency. Preprocessing instead analyzes documents once at upload time and reuses the resulting artifacts across multiple queries, substantially reducing operational cost when query frequency is high. Effective preprocessing requires that extraction goals be specified in advance; DeepVis addresses this through user-directed prompts at document upload time, guiding vision analysis toward the information most relevant to the intended research domain.

C. Document Understanding Systems

Recent vision-capable foundation models have demonstrated the ability to handle diverse document formats without task-specific training or fine-tuning, aligning with real-world research needs where document types and query patterns cannot be fully anticipated during system design. DeepVis extends this capability through user-directed prompts specified at document upload time, guiding vision analysis toward domain-relevant information without requiring specialized model architectures or document templates.

D. Multi-Step Document Processing Pipelines

Recent work on long document summarization has explored multi-step preprocessing architectures to address model context limits. Sie et al. [11] propose a multi-step extractive-abstractive pipeline for regulatory documents, where extractive summarization compresses documents before abstractive generation, finding that compression strategies interact with model architecture in ways that affect summary quality. While this approach shares DeepVis's insight that preprocessing enables more effective downstream inference, the two strategies differ fundamentally: extractive summarization operates on OCR-extracted text, discarding the visual and spatial information that DeepVis's vision-based preprocessing preserves. Extractive selection also fragments page-level coherence by selecting sentences across document sections, making accurate page citation unreliable. DeepVis's directed approach instead performs full-page vision analysis in deterministic page order, maintaining coherent document structure throughout and enabling the exact page references required for verifiable research queries.

E. Positioning DeepVis

DeepVis occupies a distinct position between chunk-based retrieval and direct vision LLM analysis. RAG systems select relevant content at query time through semantic similarity search, operating effectively on text-heavy collections with diverse queries but losing visual information and page structure through OCR-based preprocessing and non-deterministic chunk ordering. Direct vision LLM approaches preserve full information fidelity by processing raw page images at query time, but per-query inference costs scale prohibitively for large collections or high query volumes. DeepVis bridges these approaches through one-time vision preprocessing: documents are analyzed using vision-capable LLMs at upload time according to user-specified directions, and the resulting artifacts are organized in strict document and page order for efficient, deterministic inference. This design targets collections with consistently high query activity where visual understanding and verifiable citations are essential requirements.

The appropriate choice among approaches depends on document characteristics and usage patterns. RAG remains preferable for primarily text-based collections, collections that exceed context window limits, or use cases where query diversity makes preprocessing direction impractical to specify. Direct vision processing is appropriate when collections are very small or when queries are sufficiently unpredictable that no preprocessing strategy could adequately anticipate the needed analysis.

III. Methodology and System Design

A. Architecture Overview

DeepVis employs a two-stage architecture that separates preprocessing from inference, optimizing cost and performance through deliberate placement of computational operations.

This separation reflects a fundamental insight: for document collections with variable ingestion patterns but consistently high query activity, the cost structure shifts dramatically. Rather than paying operational costs for every query in RAG or processing thousands of pages for each question in direct vision LLM approaches, DeepVis incurs a one-time preprocessing cost that enables near-instant subsequent queries. The architecture trades query-time flexibility for deterministic performance, lower operational costs, and enhanced reliability through consistent context ordering.

The pipeline's modular design allows the system to adapt to evolving LLM capabilities, as demonstrated by our migration from GPT-4o to GPT-4.1, which improved performance and reduced costs without requiring reprocessing of existing document collections.

B. Preprocessing Pipeline

1) Document Upload and User Direction: DeepVis's preprocessing begins when users upload documents and specify a per-document prompt that directs the vision analysis. This user direction is critical to the system's effectiveness: rather than attempting to extract all possible information from each page, which would be both expensive and unwieldy, users specify what information matters for their research domain.

For art history material, a typical direction prompt might be: "Summarize artistic movements and techniques mentioned or shown, with detailed visual descriptions of artworks including composition, color palette, and style. Note any artist names, dates, and historical context." For administrative documents, users might specify: "Extract policy statements, tabular data, and accompanying images." This flexibility allows DeepVis to

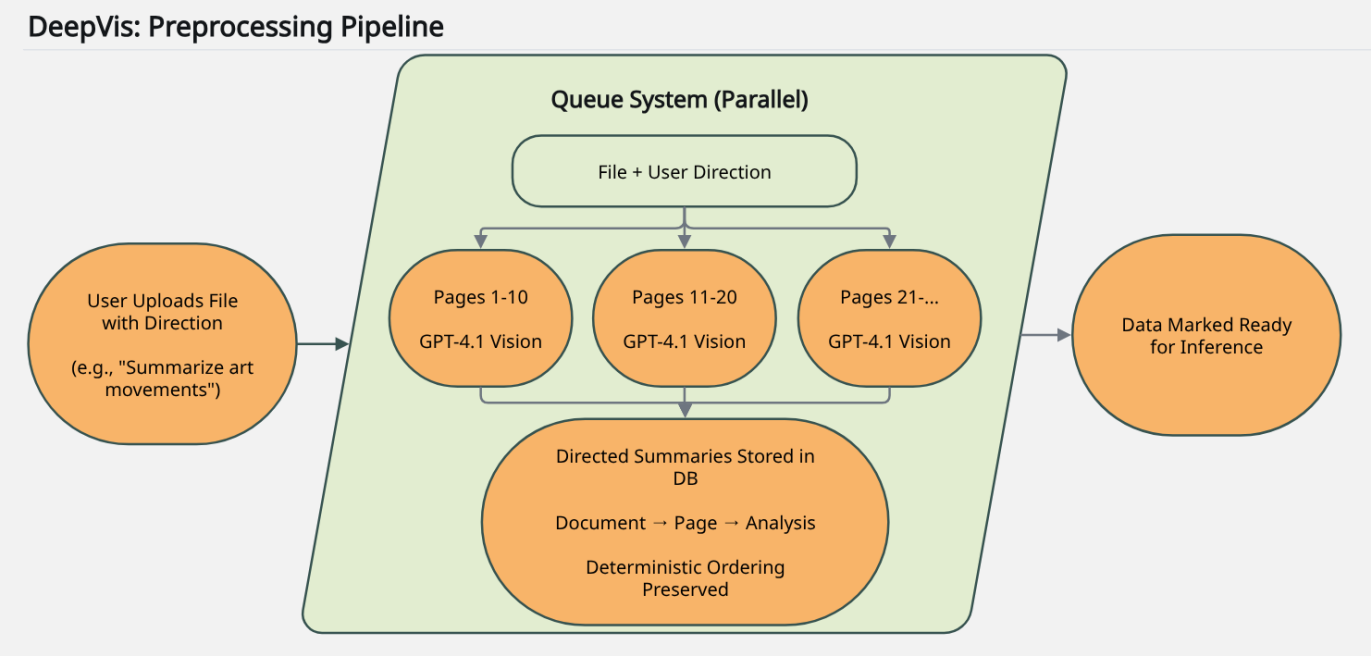

Fig. 1. Overview of stage 1 of DeepVis: User uploads a file with direction, the system processes the file in parallel using a vision-capable LLM, and stores directed analyses and extractions ready for stage 2 inference.

serve diverse use cases without modification to the underlying architecture.

The direction prompt is stored alongside the document and applied consistently to every page during preprocessing, ensuring coherent analysis across the entire document. Users can also upload the same document multiple times with different direction.

2) Page-by-Page Vision Analysis: The core of DeepVis's preprocessing pipeline processes each document page as a multimodal input combining the page image with the user's direction prompt. Vision-capable LLMs (initially GPT-4o, later GPT-4.1) analyze each page to extract textual, visual, and spatial information according to the specified direction and store them sequentially.

This vision analysis preserves information that OCR-based approaches discard or misrepresent. For art history material, the system can identify artistic movements visible in images, describe compositional techniques, note color palettes, and understand how captions relate to artworks regardless of their spatial arrangement on the page. For documents with complex layouts, the vision model comprehends how diagrams connect to surrounding text, how tables structure information, and how visual hierarchy conveys meaning. For degraded or handwritten documents, the vision approach can often interpret content that OCR systems fail to extract accurately. For example, DeepVis was able to accurately produce the text of an old handwritten music score with illegible unaligned accompanying text, demonstrating robustness beyond traditional OCR for degraded historical materials.

The output of vision analysis is a directed analysis and extraction for each page that follows the structure given by the user's direction prompt. Critically, each analysis and extraction maintains its exact page number, enabling precise citation in downstream responses. The page-by-page processing ensures that context is preserved at natural document boundaries rather than fragmented across arbitrary chunks as in RAG systems.

In our pilot implementation, preprocessing cost averages approximately $0.005 per page, a one-time expense that enables unlimited subsequent queries without embedding calls. For a 1,000-page document collection, the total preprocessing cost is approximately $5.

3) Analysis and Extraction Storage: Preprocessed analyses and extractions are stored with deterministic organization, ordered by document identifier and then by page number within each document. This seemingly simple design choice has profound implications for system reliability and LLM performance.

By maintaining document and page order, DeepVis creates a coherent information structure that mirrors how humans understand documents and how LLMs are trained on those documents: as sequential narratives. When these analyses and extractions are later provided to an LLM, the model receives information in a consistent, logical order that reflects the document's inherent structure. This contrasts sharply with RAG systems, where semantic similarity search returns chunks in non-deterministic order based on vector distances, query embeddings, and index state.

Deterministic ordering produces consistent results across repeated queries, limits hallucination risk by providing coherent context rather than disparate fragments the LLM must reconcile, and enables efficient caching since the page collection remains stable across queries.

Analyses and extractions are stored as simple ordered text artifacts, requiring no specialized vector databases or embedding infrastructure and reducing operational complexity compared to RAG alternatives.

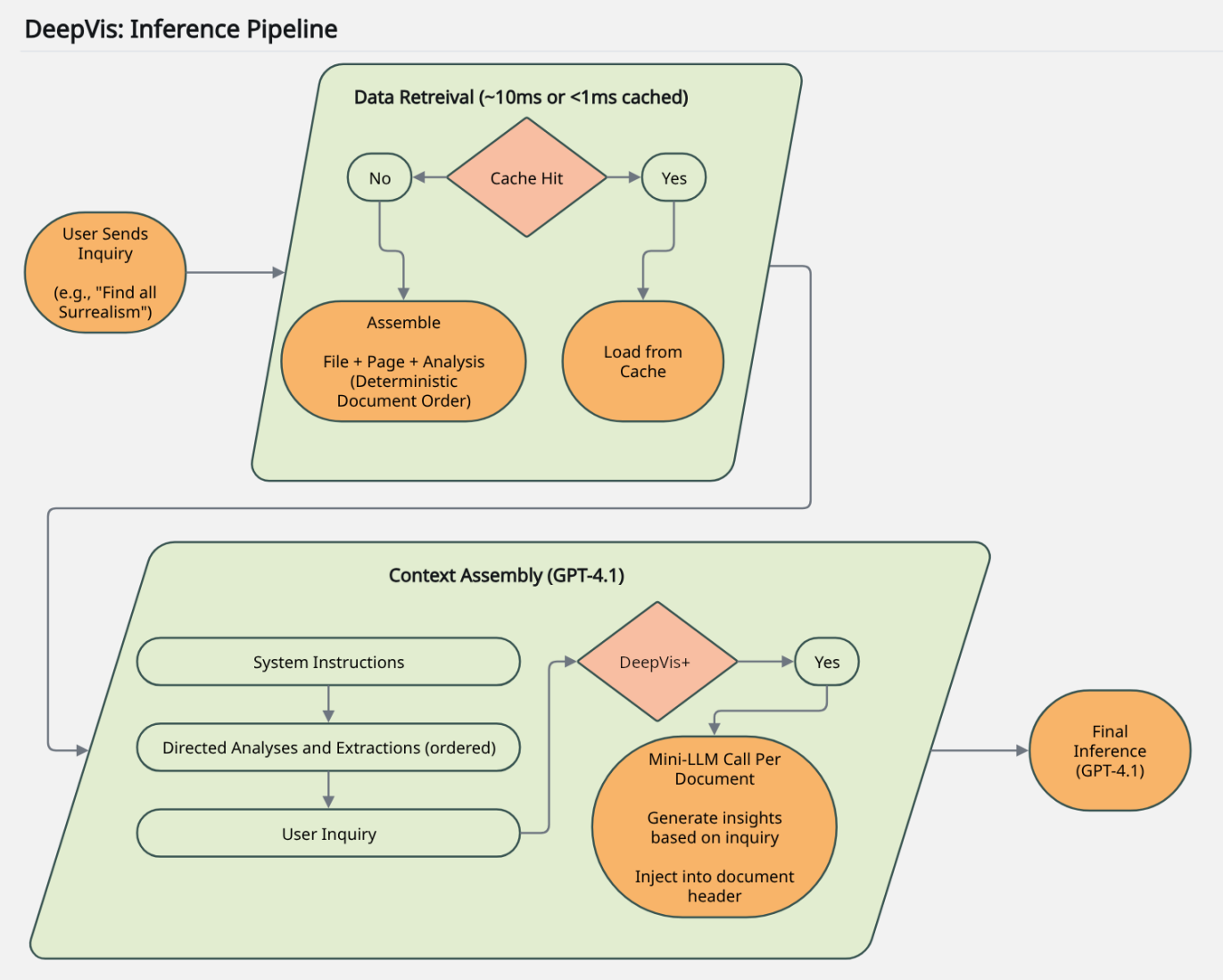

Fig. 2. Overview of stage 2 of DeepVis: User sends an inquiry, analyses and extractions are retrieved from cache or assembled, and context is assembled for the final LLM call containing system instructions, directed analyses and extractions, and the user inquiry. DeepVis+ extends this stage with a mini-LLM insight pass injected into the assembled context (see Section V-B).

C. Inquiry-Time Processing

1) Context Construction: When a user submits a research query, DeepVis constructs the LLM context through a three-component structure that strategically organizes information to maximize model effectiveness:

System Prompt: The system prompt provides instructions for query handling, citation format, and response strategy. It directs the model to cite specific page numbers when making claims, acknowledge uncertainty when information is not present in the analyses and extractions, and maintain fidelity to the preprocessed content rather than relying on parametric knowledge.

Document Analyses and Extractions: The preprocessed page-by-page analyses and extractions are inserted in strict document and page order. For a query spanning multiple documents, analyses and extractions from all relevant documents are included, maintaining their deterministic organization. Each analysis and extraction retains its page reference, enabling the LLM to cite specific pages in its responses. The deterministic ordering creates a coherent "memory" structure that the model can navigate systematically rather than attempting to synthesize information from unordered fragments.

User Prompt: The user's research question or inquiry appears at the end of the context, following standard prompt engineering practices that position the immediate task after relevant background information. This structure allows the LLM to process the organized analyses and extractions before encountering the specific query, enabling more systematic analysis. In a later iteration, DeepVis+ leverages the user prompt for document-level insight generation to further tune the input into the LLM as to which page context is relevant to the active user inquiry (see Section V-B).

This three-component structure effectively stretches the limits of available context windows by organizing information deterministically. Rather than requiring semantic search to identify relevant content, the system provides comprehensive

page-by-page coverage in a form optimized for LLM consumption.

2) LLM Inference: With context constructed, a large context window LLM processes the entire structured input to generate responses. The deterministic context ordering, combined with explicit page references throughout the analyses and extractions, enables several key capabilities:

Page-Level Precision: Because each analysis and extraction maintains its page number and the model can process the full ordered context, responses can cite specific pages with high accuracy. When answering "Find all instances of Surrealism," the system can return page numbers for each mention, enabling users to verify claims and conduct deeper investigation. This precision is critical for academic research and distinguishes DeepVis from RAG approaches where chunking and retrieval make accurate page citation difficult.

Deterministic Results: Given the same document collection and query, DeepVis produces consistent responses because the context provided to the LLM remains constant. This deterministic behavior enhances system reliability and user trust, particularly important for research applications where reproducibility and verifiability matters. In contrast, RAG systems may return different chunks in different orders depending on index state and retrieval parameters, leading to variable responses.

Comparison Across Documents: The ordered, comprehensive context enables the LLM to compare information across pages and documents systematically. For queries asking for the evolution of a particular concept, the model can trace conceptual development across multiple pages, identifying relationships and progressions.

Performance Characteristics: Query response time in our implementation averages approximately 10 milliseconds per page of summarized content when not yet cached, or under 1 millisecond per page with warm caches for repeated queries on the same document collection. This represents a significant performance advantage over RAG systems, which in our testing required 400–1000 milliseconds per query to perform semantic search, retrieve chunks, and generate responses. The combination of deterministic context and elimination of retrieval overhead produces near-instant data for the final LLM call in research workflows.

Reduced Hallucination Risk: DeepVis's deterministic context organization presents information as a coherent sequential narrative rather than unordered, potentially contradictory fragments — a condition shown to exacerbate both fact fabrication and omission in RAG systems [2]. System prompt instructions require the model to ground all assertions in specific preprocessed analyses with explicit page citations, creating structural accountability that discourages extrapolation beyond provided content. Together, these properties relocate hallucination risk from the unpredictable inference stage to the bounded preprocessing stage; observational support appears in the evaluation, where RAG's 1.2% page citation precision and the Section 4 case study both reflect fabrication patterns consistent with citation hallucination.

D. Implementation Details

DeepVis was initially developed in August 2024 using GPT-4o and later migrated to GPT-4.1 following its release in April 2025. GPT-4.1 offers a 1M token context window compared to GPT-4o's 128K, improved instruction following (38.3% on MultiChallenge), and reduced operational costs. The modular pipeline architecture enabled this migration without reprocessing existing document collections: only the inference stage required updating, while preprocessed analyses and extractions remained fully compatible. Future preprocessing runs take advantage of GPT-4.1's improved vision capabilities directly.

GPT-4o's 128K context window supported between 250 and 1,000 pages per query, with individual analyses and extractions ranging from 120 to 500 tokens. GPT-4.1's expanded context window extends this capacity to approximately 10,000 pages simultaneously, sufficient for substantial research document collections. For collections exceeding this limit, queries can be structured against relevant document subsets or aggregated across multiple inference passes; the current implementation requires users to organize these subsets manually, with automated partitioning planned in future work (see Section V-E).

Preprocessing uses parallel processing in 10-page groupings to optimize throughput while maintaining ordered output. Preprocessed analyses and extractions typically consume approximately 25% of original document size, requiring only simple ordered text storage with no specialized vector databases or embedding infrastructure.

IV. Results and Evaluation

A. Evaluation Setup

1) Corpus and Motivation: We evaluate DeepVis and RAG using Owen Jones's *The Grammar of Ornament* [12], a canonical art history reference first published in 1856. The corpus comprises 612 pages including 100 hand-lithographed color plates spanning ornamental traditions from ancient Egyptian and Greek design through Moresque, Byzantine, Celtic, and Renaissance European styles. Pages were processed as-is, including pages with degraded or damaged text resulting from age and reproduction quality of the source scan.

This corpus was selected for three reasons. First, it is multimodal by design: the ornamental plates must be seen to be analyzed, as visual pattern geometry, color palette, and compositional arrangement carry semantic meaning that OCR-extracted text cannot convey. Second, it provides verifiable ground truth anchors in the form of labeled plates and supporting text, enabling objective assessment of citation accuracy and factual correctness. Third, it directly represents DeepVis's original art history deployment context, providing an ecologically valid evaluation environment.

2) System Configuration: Table I summarizes the configuration for each system. Both systems received queries through the same ZotGPT platform interface, establishing a consistent comparison baseline. The shared user query prompt references "image and document data" as part of the ZotGPT platform's general-purpose inference layer; at inference time, DeepVis supplies preprocessed text analyses and extractions rather than raw page images.

TABLE I
EVALUATION SYSTEM CONFIGURATION

Parameter	Configuration
DeepVis Preprocessing LLM	GPT-4.1, image detail: high
DeepVis Preprocessing Prompt	General-purpose visual description: capture text verbatim including structure and layout; describe visual color, geometry, patterns, and motifs; note relationships between visuals and text
RAG Processing	Azure Cognitive Search, model `prebuilt-read`, paragraph chunking, embeddings via `text-embedding-3-small`
Query Interface	ZotGPT ClassChat (identical for both systems)
Text Accuracy Tolerance	Verbatim match with allowance for damaged source material: up to 1 missed word and up to 3 total characters

DeepVis preprocessing used a general-purpose visual description prompt rather than a domain-specific ornamental art direction. The choice to configure preprocessing with general directions rather than specialized academic prompts helps demonstrate that DeepVis's architectural advantages derive from the preprocessing and deterministic organization strategy rather than from domain-tailored prompting, even though tailored prompting would yield better results. As discussed in Section IV-B2, this choice is the most direct explanation for Category M underperformance and represents the clearest path to improvement.

3) Query Design: We constructed 15 evaluation queries across three categories of five queries each, stratified by modality requirements.

Category T — Text-Accessible (Control): Queries answerable from text content alone, where both systems are expected to perform comparably. This category establishes that DeepVis does not regress on standard text retrieval and provides a performance baseline for interpreting differences in other categories.

Category V — Visually-Dependent (Key Differentiator): Queries requiring analysis of ornamental plate content, including identification of geometric patterns, color palettes, motif types, and compositional arrangements. OCR-based systems are expected to fail because the relevant information resides in visual content that text extraction cannot capture. This category provides the primary evidence for DeepVis's multimodal advantage.

Category M — Cross-Modal Synthesis: Queries requiring integration of textual chapter content with visual plate analysis, such as verifying whether plates and associated caption text can be visually located. This category tests the full multimodal pipeline capability and places the greatest demand on preprocessing completeness.

4) Ground Truth and Annotation Protocol: Ground truth was established prior to system evaluation by a single annotator following a source-anchored protocol designed to minimize subjectivity. For each category, each query answer page was decided before running any system queries. For Category T, ground truth was derived directly from printed text. For Category V, ground truth was derived using simple visual observations based on direct plate observation. For Category M, ground truth was derived combining textual and visual evidence as in Categories T and V.

B. Quantitative Results

Table II presents measured metrics across both systems. Results are reported as counts alongside percentages given the per-category sample size of five queries; aggregate trends across categories provide the primary interpretive basis.

1) Query Success Rate: Category T results confirm that DeepVis does not regress on text retrieval relative to RAG, with both systems achieving 4/5 (80%). DeepVis's single failure in this category occurred when preprocessing did not fully capture verbatim text from a page despite prompt instructions, leaving insufficient information for an accurate response. RAG's single failure reflected its lack of page boundary awareness: the system returned large page ranges in an attempt to cover the answer area rather than identifying a specific location, consistent with chunk-based retrieval behavior where document structure is not preserved.

Category V results reveal the central functional difference between the two systems. DeepVis achieved 4/5 (80%) compared to RAG's 2/5 (40%), but this comparison understates the gap in visual capability. As noted in Table II, both of RAG's correct Category V responses were derived from supporting text co-located with plates in the document rather than from any visual analysis of plate content. RAG demonstrated no ability to identify ornamental features, color relationships, or compositional patterns from visual content; its correct answers were coincidental byproducts of text proximity rather than visual understanding. DeepVis's Category V responses consistently referenced specific plate characteristics observed during preprocessing, including motif geometry, palette composition, and spatial arrangement unavailable to OCR-based systems.

2) Category M and Static Preprocessing: Category M results highlight the current limitation of DeepVis's static preprocessing approach. DeepVis achieved 2/5 (40%) compared to RAG's 0/5 (0%). Failures in both systems were attributable in part to damaged source material that the processing pipeline could not accurately parse; one DeepVis failure fell within the defined text accuracy tolerance, reflecting partial success under challenging source conditions. The use of a general-purpose preprocessing prompt, rather than a domain-specific direction emphasizing object orientation, count and design principles, contributed to the remaining Category M failures: analyses and extractions did not consistently capture the visual context needed for cross-modal synthesis queries. A domain-specific preprocessing prompt addressing these areas would be expected to substantially improve Category M performance.

3) Page Citation Precision and Recall: Page citation results provide the starkest system differentiation in the evaluation. DeepVis achieved 100% PCR, correctly identifying the source page for every one of the 15 queries. RAG achieved 6.7% PCR (1/15); the single correct citation was attributable to a scanned page number embedded within a retrieved text chunk rather than any structural page awareness in the system.

The PCP denominators reflect meaningfully different citation behaviors: DeepVis cited an average of 1.6 pages per

TABLE II
DEEPVIS VS. RAG EVALUATION ON *The Grammar of Ornament*

Metric	Category	DeepVis	RAG
QSR	T: Text-Accessible — Control Set	4/5 (80%)	4/5 (80%)
QSR	V: Visually-Dependent — Key Differentiator	4/5 (80%)	2/5 (40%)†
QSR	M: Cross-Modal Synthesis	2/5 (40%)	0/5 (0%)
Total QSR	T, V, and M	10/15 (67%)	6/15 (40%)
PCP	T, V, and M	15/24 (62.5%)	1/81 (1.2%)
PCR	T, V, and M	15/15 (100%)	1/15 (6.7%)
VCA	V only	4/5 (80%)	0/5 (0%)

† Both RAG Category V correct answers were derived from supporting text co-located with plates, not from visual analysis. See Section IV-B4.

query (24 total citations across 15 queries; citations beyond the primary answer page reflect topically related content), while RAG cited an average of 5.4 pages per query (81 total citations), achieving 1.2% PCP (1/81). RAG's high citation volume with near-zero accuracy reflects a characteristic failure mode of chunk-based retrieval without page metadata: the system defaulted to citing early document pages, particularly page 1, likely reflecting chunk retrieval order rather than location relevance.

While custom chunking strategies with embedded page metadata could partially improve RAG citation accuracy for PDF page numbers, such improvements would not address scanned page number capture nor provide any visual understanding for Category V queries.

4) Visual Content Accuracy: VCA results provide direct evidence for DeepVis's multimodal capability. DeepVis achieved 80% VCA (4/5), correctly identifying the relevant plates and their visual characteristics for four of five visually-dependent queries. The single failure occurred when preprocessing did not accurately identify a count of leaf elements within a botanical motif, a fine-grained quantitative visual task that the LLM did not record during preprocessing analysis. This specific failure case also motivates DeepVis++: a mini-LLM vision pass tuned to the active user inquiry could flag leaf-count details as relevant before final inference, improving precision on queries requiring fine visual enumeration.

RAG achieved 0% VCA (0/5). Responses to visually-dependent queries demonstrated exclusively text-based reasoning with no reference to actual plate imagery. When asked to identify plates by visual characteristics, RAG was unable to specify plate locations and instead provided speculative guidance on where relevant images might be found within the document, confirming that OCR-based preprocessing offers no path to visual content retrieval. The 80 percentage point gap between DeepVis and RAG on VCA is the core multimodal capability result of this evaluation.

C. Qualitative Case Studies

1) Category V Success: Visual Color Identification in EGYPTIAN N°3: Query V2 asked for the dominant color of the largest Egyptian column capital in "EGYPTIAN N°3." The correct answer is red, found on page 59 — a fact that requires direct visual inspection of the plate and cannot be reliably inferred from surrounding text.

DeepVis returned the correct answer with an exact page citation: *"The dominant color… is **red**… page 59… decorated with repeating pointed leaf patterns in red, yellow, and green, but red is the most visually dominant color on this capital."* RAG returned blue, citing pages 10, 18, and 30, having applied Jones's general textual principle that "blue always occupies the largest area" as though it described the specific plate.

RAG's failure illustrates a systematic risk beyond simple inaccuracy: the system constructed a confidently formatted response, including a summary table and multi-point citation list, from general color theory rather than visual plate content. Page 59 was not cited. DeepVis's preprocessing captured the observed color distribution during the preprocessing stage, grounding the response in plate content rather than textual generalization.

2) Category M Partial Failure: Static Preprocessing and Fine-Grained Spatial Detail: Query M1 asked for the facing directions of the first Egyptian and Assyrian illustrations in the Assyrian and Persian Ornament chapter. The correct answer is Egyptian (top) right, Assyrian (bottom) left, both on page 86. No text in the document states the facing direction of either figure; the answer requires locating the page by chapter context and reading orientation from the image.

DeepVis located page 86 and returned: *"The upper illustration… facing **to the right**… The lower illustration… also facing **to the right**. Page: 86."* The Egyptian direction and page citation are correct; the Assyrian direction was recorded incorrectly during preprocessing. RAG cited pages 3–26 and described all illustrations as "upright, front-facing" from general decorative plate conventions, failing to locate page 86 or assess directional orientation for either figure.

DeepVis's failure is narrow and architectural: the spatial detail was not captured during the general-purpose preprocessing pass and cannot be corrected by reframing the query, because the error is fixed in the preprocessing artifact before any query is issued. This directly motivates DeepVis++, which would retrieve the source image of page 86 on demand and perform targeted re-analysis of the Assyrian chariot's orientation before generating the final response (see Section V-B). RAG's failure, by contrast, is categorical: without image access, the system cannot reason about directional orientation at all.

D. Pilot Deployment Findings

In addition to the formal evaluation, DeepVis was piloted from September through November 2024 using an art history corpus comprising thousands of pages of visually-rich material, assessing real-world research behavior over a full academic quarter.

The pilot involved Dean Tyrus Miller's graduate-level Visual Studies 295 seminar at UCI, where the curated corpus demonstrated value in providing students access to relevant information across advanced texts with greater accuracy and course relevance than general AI chatbots [13]. Critical feedback identified static preprocessing as the primary gap: graduate-level inquiries fell outside the scope of general-purpose analyses and extractions, directly paralleling the Category M findings of the formal evaluation. The development team responded by adding experimental capabilities for more context-aware responses, described in Sections V-B and beyond.

The pilot confirmed three core strengths of the system. Deterministic results built user trust where faculty had previously experienced variable RAG responses to identical queries. Accurate page citations supported research verification workflows, with users reporting increased confidence in system outputs for activities where accuracy mattered. The system's ability to handle visually-rich and degraded content that OCR-based approaches had misrepresented or missed expanded the range of materials researchers could effectively query.

Following the pilot, DeepVis expanded beyond art history to administrative assistants and public-facing chatbots across UCI, with students encountering the system in classrooms starting Winter 2025 [13].

V. Limitations and Future Work

A. Current Limitations

1) Quote and Verbatim Content Accuracy: DeepVis's preprocessing compresses page content according to user-directed prompts, trading verbatim accuracy for efficiency and context window utilization. Exact wording is preserved only when explicitly directed during initial processing. For research requiring precise quotations, users must verify specific wording by consulting original page images using the provided page citations.

2) Static Preprocessing: The preprocessing architecture creates analyses and extractions based on user-directed prompts specified at document upload time, but these artifacts remain static once generated. User inquiries at query time do not influence the content of preprocessed analyses and extractions, meaning the system can only surface information that was extracted during initial processing. When users pose questions requiring details not captured by the initial direction prompt, the preprocessed analyses and extractions may lack relevant information even if it exists in the source documents. In this situation, the architectural properties that limit hallucination risk — coherent sequential context and explicit page citations grounding model assertions — are also compromised: incomplete preprocessing context increases the likelihood that the model will extrapolate beyond provided content to fill gaps.

3) Modality Constraints: The current DeepVis implementation processes only image and text modalities, leveraging vision-capable LLMs to analyze document pages containing visual and textual information. The system's architecture is tied to vision LLM capabilities, and expansion to alternative modalities would require replacing the vision model with LLMs capable of processing those specific input types. While the fundamental preprocessing philosophy of creating directed artifacts in deterministic order should transfer to other modalities, practical implementation for audio or video content has not yet been tested. This modality constraint limits DeepVis's applicability to document-based materials; planned multimedia extension work is described in Section V-D.

B. Active Development: DeepVis+

DeepVis+ adds a query-responsive insight layer between preprocessing and final inference. A mini-LLM processes the preprocessed analyses and extractions in the context of the active user inquiry, generating document-level insights that identify which pages and content are most pertinent and injecting this enhanced context into the assembled prompt. This bridges the gap between static preprocessing and dynamic user questions without requiring full reanalysis of source documents, improving response relevance while preserving the deterministic organization and cost-effectiveness of the base system. The intermediate analysis layer adds latency compared to the base system but substantially improves adaptability to diverse user needs.

C. Planned Work: DeepVis++

DeepVis++ introduces agentic page retrieval for on-demand fine-grained visual analysis. When the inference model determines that preprocessed analyses and extractions lack sufficient detail to answer a query, an agent LLM retrieves specific source page images for targeted re-analysis before generating the final response, combining the cost-effectiveness of preprocessing for broad queries with the fidelity of real-time vision analysis for targeted deep dives. The Category M partial failure documented in Section IV-C2, where a spatial orientation detail was not recorded during general-purpose preprocessing, illustrates the exact class of error this architecture is designed to address.

D. Planned Work: Multimedia Extension

Building on DeepVis's preprocessing philosophy, planned work extends the architecture to audio and video modalities. Lecture recordings, podcasts, and video presentations would be preprocessed into directed artifacts with temporal citations, applying the same one-time analysis cost and deterministic organization strategy used for documents. The primary implementation change is substitution of the vision LLM with multimodal models capable of processing audio and video inputs; the preprocessing pipeline, deterministic ordering, and inference stage require no architectural modification. This cost advantage is especially compelling for audio and video, where real-time LLM inference is substantially more expensive per query than for documents.

E. Unexplored Directions

Adaptive summarization offers a longer-term improvement path: by analyzing user query patterns to identify gaps between requested and captured information, the system could suggest refined preprocessing prompts or flag documents for targeted reprocessing, progressively improving preprocessing quality from observed usage without requiring architectural changes.

VI. Conclusion

Research across large collections of visually-rich documents has historically challenged automated systems, with Retrieval-Augmented Generation approaches proving inadequate for use cases requiring visual understanding, precise citations, and verifiable results. RAG systems' reliance on OCR-based text extraction discards critical visual and spatial information, while their semantic chunking produces non-deterministic, fragmented context that obscures page-level structure and increases hallucination risk. These limitations became apparent when art history researchers sought to analyze thousands of pages of material with queries like "Find all instances of Surrealism," where existing RAG implementations failed to provide either the visual comprehension or accurate page references necessary for academic work.

DeepVis addresses these challenges through three key innovations. First, vision-based preprocessing leverages multimodal LLMs to analyze documents page-by-page, generating user-directed analyses and extractions that preserve textual, visual, and spatial information according to prompts specified at upload time. Second, deterministic organization arranges these analyses and extractions in strict document and page order within the LLM's context space, creating coherent memory structures that limit hallucination risk and enable consistent, reproducible results. Third, exact page references maintained throughout preprocessing and inference allow users to verify system responses and conduct deeper investigation with confidence.

The system's progression from initial prototype to university-wide deployment demonstrates both its technical effectiveness and practical value. DeepVis successfully answered complex research queries during pilot testing from September through November 2024 that RAG implementations could not support. Following critical and constructive feedback, the system expanded beyond art history to administrative assistants and public-facing chatbots across UCI, validating the generality of the directed preprocessing approach.

The one-time preprocessing design, deterministic organization, and cached inference yield architectural cost and performance properties that favor high-query deployments over both large-scale RAG systems and per-query real-time vision approaches, persisting as the system scales. The active development roadmap addresses the primary limitation identified in evaluation: DeepVis+ injects query-responsive insights into the assembled context, and DeepVis++ will agentically retrieve source material when static preprocessing proves insufficient.

More broadly, DeepVis suggests a reconsideration of the retrieval-versus-preprocessing trade-off in LLM knowledge systems. While RAG's semantic retrieval excels for text-heavy collections with diverse, unpredictable queries, preprocessing strategies offer compelling advantages for use cases with variable document ingestion patterns but consistently high query activity against selected collections, particularly when visual understanding, verifiability, and deterministic results matter. DeepVis's successful deployment demonstrates that for research and chatbot applications requiring reliable, verifiable responses over visually-rich materials, directed preprocessing with deterministic organization provides capabilities that current Retrieval-Augmented Generation architectures fundamentally cannot match.

Acknowledgments

The author thanks Dean Tyrus Miller and his seminar students for pilot feedback informing system refinements, and the UC Irvine ZotGPT team for testing, UI/UX design, and user research driving future development. Thanks also to the UC Irvine Office of Information Technology for supporting DeepVis development within the ZotGPT platform.

References

[1] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," 2024, arXiv:2312.10997.

[2] Z. Ma, S. An, Z. Lin, Y. Zou, J.-G. Lou, and B. Xie, "Enhancing Retrieval-Augmented Generation with Dehallucinating Parallel Context Extension," in Proc. LCFM Workshop, Int. Conf. Mach. Learn. (ICML), Vancouver, BC, Canada, Jul. 2025. [Online]. Available: https://openreview.net/forum?id=iSS4ufmtB8

[3] Microsoft, "Retrieval Augmented Generation (RAG) in Azure AI Search," *Microsoft Learn*, Apr. 22, 2024. Accessed: Jun. 7, 2024. [Online]. Available: https://web.archive.org/web/20240607113900/https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview

[4] P. Achkar, T. Gollub, and M. Potthast, "Ask, Retrieve, Summarize: A Modular Pipeline for Scientific Literature Summarization," in *Proc. SCOLIA 2025, First Int. Workshop Scholarly Inf. Access, ECIR 2025*, Lucca, Italy, 2025, pp. 41–56. [Online]. Available: http://ceur-ws.org/Vol-4022/#paper-05

[5] OpenAI, "GPT-4V(ision) System Card," San Francisco, CA, USA, Tech. Rep., Sep. 2023. Accessed: Dec. 23, 2025. [Online]. Available: https://cdn.openai.com/papers/GPTV_System_Card.pdf

[6] OpenAI, "GPT-4o System Card," 2024, arXiv:2410.21276.

[7] OpenAI, "Introducing GPT-4.1 in the API," *openai.com*, Apr. 14, 2025. Accessed: Dec. 23, 2025. [Online]. Available: https://openai.com/index/gpt-4-1/

[8] T. A. Tran, K. Oh, I.-S. Na, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "A robust system for document layout analysis using multilevel homogeneity structure," *Expert Syst. Appl.*, vol. 85, pp. 99–113, 2017, doi: 10.1016/j.eswa.2017.05.030.

[9] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Virtual Event, CA, USA, 2020, pp. 1192–1200, doi: 10.1145/3394486.3403172.

[10] D.-L. Li, S.-K. Lee, and Y.-T. Liu, "Printed document layout analysis and optical character recognition system based on deep learning," *Sci. Rep.*, vol. 15, Art. no. 23761, 2025, doi: 10.1038/s41598-025-07439-y.

[11] M. Sie, R. Beek, M. Bots, S. Brinkkemper, and A. Gatt, "Summarizing Long Regulatory Documents with a Multi-Step Pipeline," in *Proc. Nat. Legal Lang. Process. Workshop 2024*, Miami, FL, USA, Nov. 2024, pp. 18–32, doi: 10.18653/v1/2024.nllp-1.2.

[12] O. Jones, *The Grammar of Ornament*. London, U.K.: Day and Son, 1856. [Online]. Available: https://archive.org/download/gri_33125008700086/gri_33125008700086.pdf

[13] J. Nguyen, "Classroom-focused AI chatbot is making its way into UCI classrooms," *UC Irvine Office of Information Technology*, Jan. 28, 2025. [Online]. Available: https://www.oit.uci.edu/2025/01/28/classchat-story/

Christopher Robert Price is an AI Architect and Lead AI Developer at the University of California, Irvine, Office of Information Technology. He specializes in software application architecture and AI systems development, with a research focus on the design and practical deployment of novel software solutions at the intersection of emerging technology and unexplored problem spaces. His research interests include multimodal document understanding, vision-language systems, and computer security. He holds degrees in Computer Science and Japanese, and has extensive experience building enterprise software systems including the ZotGPT platform.