Foundations

Multimodal Fundamentals

PDFs, Layout, Tables/Charts, Images, and Audio

Michael J Bommarito II · Jillian Bommarito · Daniel Martin Katz

December 27, 2025

Working Draft Chapter

Version 0.1

Practical multimodal workflows for legal and finance: preserve structure, extract tables, summarize audio, and protect privacy.

Contents

How to Read This Chapter **4**

0.1 Introduction and Scope. **4**

0.1.1 The Multimodal Imperative 4

0.1.2 Building on Prior Foundations. 5

0.2 Images and Visual Content. **5**

0.2.1 Image Understanding in Legal and Financial Context 5

0.2.2 OCR: From Scanned Documents to Text 6

0.2.3 Handwriting and Signature Recognition 7

0.2.4 Diagrams, Flowcharts, and Schematics. 8

0.2.5 Image Classification and Document Routing 9

0.2.6 Screenshot and UI Analysis 9

0.2.7 Visual PII and Privilege Markers. 10

0.3 Document Structure and Layout **11**

0.3.1 The Spectrum of Document Parsing Strategies 11

0.3.2 Layout Analysis Models 11

0.3.3 Preserving Structure for Downstream Use 12

0.3.4 Chunking Strategies for Retrieval 12

0.3.5 OCR Quality and Preprocessing 13

0.3.6 Form and Template Extraction 13

0.3.7 Complex Layout Challenges 14

0.4 Tables and Charts **16**

0.4.1 Table Extraction Strategies 16

0.4.2 Chain-of-Table Reasoning. 16

0.4.3 Chart Understanding 17

0.4.4 Multimodal Embeddings for Tables and Charts . . . 18

0.5 Audio and Transcripts . . . 18

0.5.1 Audio RAG Pipelines . . . 18
0.5.2 Video Understanding and Retrieval . . . 19
0.5.3 Practical Considerations . . . 20
0.5.4 ASR Error Handling and Quality Control . . . 21
0.5.5 Video Frame Analysis . . . 22
0.5.6 Slide Deck Extraction . . . 23
0.5.7 Real-Time vs. Batch Processing . . . 23

0.6 Privacy and Redaction . . . 24

0.6.1 PII Detection and Redaction . . . 24
0.6.2 Privilege and Confidentiality . . . 25
0.6.3 Content Authenticity and Provenance . . . 26
0.6.4 Redaction Governance . . . 26
0.6.5 Visual Privacy Patterns . . . 27
0.6.6 MNPI Detection Workflows . . . 28
0.6.7 Cross-Border Data Considerations . . . 29

0.7 Synthesis . . . 30

0.7.1 Core Technical Themes . . . 30
0.7.2 Key Takeaways . . . 31
0.7.3 Integration Patterns . . . 31
0.7.4 Architectural Decisions . . . 32
0.7.5 What This Chapter Did Not Cover . . . 32
0.7.6 Connecting to Other Chapters . . . 33
0.7.7 Looking Forward . . . 33

0.8 Further Learning . . . 33

0.8.1 Document Layout and Structure . . . 33
0.8.2 Table and Chart Understanding . . . 34
0.8.3 Audio and Video Processing . . . 35
0.8.4 Multimodal Embeddings . . . 35
0.8.5 Privacy and Content Authenticity . . . 36
0.8.6 Practical Guides and Documentation . . . 36
0.8.7 Common Misconceptions . . . 36

0.8.8 Exercises for Practitioners 37

Conclusion 38

How to Read This Chapter

Focus on the sections aligned to your data: PDFs and layout for filings/contracts, tables/charts for financials, audio/transcripts for calls.

Key Objectives

- Preserve document structure and identifiers (pages, figures, tables).
- Extract tables/charts accurately and capture units/footnotes.
- Summarize audio with timestamps and speaker labels.
- Apply privacy/redaction safeguards before external processing.

0.1 Introduction and Scope

Retrieval-augmented generation has evolved from simple text chunking to sophisticated **multimodal RAG** pipelines capable of processing complex documents, images, and rich media. The ‘‘text-only’’ RAG pipeline is increasingly seen as legacy architecture, insufficient for enterprise data locked in PDFs, charts, slide decks, and recorded proceedings.

This chapter addresses practical multimodal inputs common to legal and finance: preserving structure in PDFs, extracting tables and charts, interpreting images and screenshots, and summarizing audio and video. We also highlight privacy and redaction patterns that must be applied before sensitive content enters AI systems.

0.1.1 The Multimodal Imperative

Legal and financial workflows are inherently multimodal:

- **SEC filings**: Combine narrative text with financial tables, charts, and embedded images.
- **Contracts**: May include scanned signatures, attached exhibits, and referenced schedules.
- **Litigation materials**: Span depositions (audio/video), exhibits (images, documents), and transcripts.
- **Regulatory correspondence**: Often arrives as scanned PDFs requiring OCR.

- **Research reports**: Integrate text analysis with data visualizations.

A system that can only process plain text misses critical information encoded in these other modalities---and may produce incomplete or misleading analysis.

0.1.2 Building on Prior Foundations

This chapter builds directly on concepts from earlier chapters:

- **Embeddings** (Chapter 1): Vector representations extend beyond text to images, tables, and audio transcripts. Models like CLIP project multiple modalities into shared embedding spaces.
- **Structured outputs** (Chapter 3): Extracted tables and metadata should conform to defined schemas for downstream integration.
- **Tool use** (Chapter 3): Document parsers, OCR engines, and ASR models act as tools that preprocessing agents invoke.
- **Evidence records** (Chapter 3): Multimodal processing steps---redaction, extraction, transcription---must be logged with the same rigor as LLM inference.

Chapter Scope

We focus on *ingestion and preprocessing*---getting multimodal content into a form suitable for LLM processing and retrieval. Generation of multimodal outputs (images, audio synthesis) is beyond our current scope, though content authenticity standards apply to both directions.

0.2 Images and Visual Content

Beyond structured documents, legal and financial workflows frequently encounter standalone images: scanned forms, photographs of evidence, screenshots of applications, diagrams, and handwritten notes. This section addresses image understanding, OCR for text extraction, and specialized visual content types.

0.2.1 Image Understanding in Legal and Financial Context

Images in professional settings differ from general-purpose computer vision tasks. The content is often text-heavy, formally structured, and carries legal or evidentiary significance:

- **Evidence photographs**: Crime scenes, property damage, product defects.
- **Scanned historical documents**: Pre-digital records, handwritten ledgers, typed correspondence.
- **Application screenshots**: UI states as evidence in software disputes, compliance monitoring.

- **Whiteboards and notes**: Meeting artifacts, informal agreements, expert sketches.
- **Identity documents**: Passports, driver's licenses, corporate filings with seals.

Vision-language models (VLMs) like GPT-4V, Claude 3, and Gemini can interpret these images directly, providing descriptions, extracting visible text, and answering questions about content. However, for systematic processing at scale, specialized pipelines remain more reliable and cost-effective.

0.2.2 OCR: From Scanned Documents to Text

Optical Character Recognition (OCR) converts images of text into machine-readable characters. For legal and financial documents, OCR quality directly impacts downstream accuracy.

OCR Engines.. Modern OCR options span a range of capabilities:

OCR Engine Comparison

Tesseract (Open Source) Mature, widely used, supports 100+ languages. Accuracy varies with image quality; benefits from preprocessing. No cost for processing.

Google Cloud Vision OCR High accuracy, handles complex layouts, returns confidence scores. Requires API access and incurs per-page costs.

Azure AI Document Intelligence Combines OCR with layout analysis and field extraction. Pretrained models for invoices, receipts, and IDs. Enterprise pricing.

Amazon Textract OCR with table and form extraction. Integrates with AWS ecosystem.

ABBYY FineReader Commercial solution with high accuracy on complex documents. Often used in legal e-discovery.

Accuracy Metrics.. OCR quality is measured by:

- **Character Error Rate (CER)**: Percentage of characters incorrectly recognized.
- **Word Error Rate (WER)**: Percentage of words with any error.
- **Confidence scores**: Per-character or per-word confidence from the OCR engine.

For legal documents, even 99% accuracy means errors in 1 of every 100 words---potentially changing the meaning of a contract clause. High-stakes applications require human review or multiple OCR passes.

Handling Low-Quality Scans.. Scanned documents often suffer from:

- **Low resolution**: Aim for 300 DPI minimum; 600 DPI for fine print.
- **Skew and rotation**: Apply deskewing algorithms before OCR.
- **Noise and artifacts**: Photocopier marks, fax degradation, coffee stains.
- **Poor contrast**: Faded ink, colored backgrounds, highlighted text.

OCR Preprocessing Pipeline

1. **Deskew**: Correct rotation and perspective distortion.
2. **Denoise**: Remove speckles, lines, and artifacts.
3. **Binarize**: Convert to black and white with adaptive thresholding.
4. **Resize**: Scale to optimal resolution for the OCR engine.
5. **Segment**: Identify text regions versus images, tables, and margins.

0.2.3 Handwriting and Signature Recognition

Handwritten content presents additional challenges beyond printed text.

Handwritten Text Recognition (HTR).. HTR models are trained specifically on handwriting samples:

- **Google Cloud Vision**: Supports handwriting detection with the `DOCUMENT_TEXT_DETECTION` feature.
- **Microsoft Azure**: Handwriting recognition through Document Intelligence.
- **AWS Textract**: Detects handwriting in form fields.
- **Transkribus**: Specialized platform for historical handwritten documents.

Accuracy varies significantly with handwriting legibility. Cursive, stylized, or hurried writing degrades recognition. For legal applications, handwritten amendments to contracts or handwritten wills require careful verification.

Signature Verification.. Signatures serve as identity verification and consent indicators. Processing involves:

- **Detection**: Locate signature regions within documents.
- **Extraction**: Isolate the signature image for analysis.
- **Comparison**: Match against known signature samples (for verification).

- **Authenticity assessment**: Detect potential forgeries or alterations.

Legal Admissibility of Signature Analysis

Automated signature verification is not universally accepted as evidence of authenticity. Courts may require expert testimony to validate AI-based signature analysis. Document the methodology, confidence levels, and limitations when signature verification is legally significant.

0.2.4 Diagrams, Flowcharts, and Schematics

Technical documents often contain diagrams that encode important information:

Common Diagram Types..

- **Organizational charts**: Corporate structure, reporting relationships.
- **Process flowcharts**: Workflow diagrams, decision trees.
- **Technical schematics**: Engineering drawings, circuit diagrams, architectural plans.
- **Patent figures**: Invention illustrations with labeled components.
- **Network diagrams**: System architecture, data flow.

Extraction Approaches.. Diagram understanding combines visual and structural analysis:

- **VLM description**: Ask a vision-language model to describe the diagram, identify entities, and explain relationships.
- **Object detection**: Identify boxes, arrows, and connectors as distinct elements.
- **Graph extraction**: Convert visual structure to a machine-readable graph (nodes and edges).
- **Label extraction**: OCR the text labels and associate them with visual elements.

Patent Diagram Analysis

Patent filings typically include multiple figures with numbered components referenced in the claims. A patent analysis system should:

1. Identify figure numbers and their bounding regions.
2. Extract component labels (‘‘102’’, ‘‘104a’’, etc.).
3. Link labels to their positions in the diagram.

4. Cross-reference with claim text mentioning those components.

0.2.5 Image Classification and Document Routing

Before detailed processing, images must be classified and routed to appropriate pipelines.

Document Type Classification.. Classify incoming images by document type:

- **Invoices, receipts, purchase orders**: Route to accounts payable extraction.
- **Contracts and agreements**: Route to legal review and clause extraction.
- **Identity documents**: Route to KYC/AML verification.
- **Correspondence**: Route to communication logging.
- **Evidence photographs**: Route to case file with metadata capture.

Bates Stamps and Exhibit Labels.. Legal documents in discovery often carry:

- **Bates numbers**: Sequential identifiers for tracking pages across a case.
- **Exhibit stamps**: Labels indicating exhibit designation (‘‘Plaintiff’s Exhibit 1’’).
- **Confidentiality markings**: Designations like ‘‘Confidential’’ or ‘‘Attorneys’ Eyes Only.’’

Automated detection of these markings enables routing, access control, and citation generation.

0.2.6 Screenshot and UI Analysis

Application screenshots serve as evidence in software disputes, compliance monitoring, and user research.

UI Element Detection.. Specialized models can identify:

- **Buttons, menus, and controls**: What actions were available.
- **Text fields and their content**: What data was displayed or entered.
- **Error messages and notifications**: System state at capture time.
- **Layout and visual hierarchy**: How information was presented.

Temporal Context.. Screenshots capture a moment in time. Metadata should include:

- Timestamp of capture (if available from system metadata).
- Application and version being captured.

- User context (if relevant and permissible).
- Screen resolution and display settings.

For compliance monitoring (e.g., trading desk surveillance), screenshots may be captured systematically and require automated analysis at scale.

0.2.7 Visual PII and Privilege Markers

Images may contain sensitive information that text-based PII detection misses.

Faces and Biometric Data.. Photographs containing faces require special handling:

- **Detection**: Identify face regions using face detection models.
- **Blurring/redaction**: Apply blur or solid overlay to protect identity.
- **Consent tracking**: Document whether face capture was consented.
- **Biometric regulations**: BIPA (Illinois), GDPR, and other laws restrict biometric data processing.

Signatures and Handwriting.. Handwritten content may reveal identity:

- Signatures are personally identifying and should be redacted in shared datasets.
- Handwriting style can be identifying; consider redaction for anonymization.

Privilege Stamps and Watermarks.. Visual markers indicating privilege status:

- **"Privileged and Confidential" stamps**: Route to secure handling.
- **"Draft" watermarks**: Indicate non-final status.
- **Firm letterheads**: May indicate attorney-client communication.

Visual Content Privacy Checklist

- Detect and handle faces per applicable biometric laws.
- Identify and redact signatures when anonymization is required.
- Recognize privilege stamps and route to appropriate controls.
- Strip or preserve EXIF metadata based on use case.
- Apply C2PA content credentials for provenance tracking.

0.3 Document Structure and Layout

The primary bottleneck in enterprise document workflows is the Portable Document Format (PDF). Standard text extraction tools treat a PDF as a stream of text characters, often destroying the semantic structure of tables, multi-column layouts, and headers. This ''PDF problem'' results in systems that can retrieve text but cannot understand the relationship between a data point in a table cell and its row/column headers---a critical failure mode for legal and financial analysis.

0.3.1 The Spectrum of Document Parsing Strategies

Modern document parsing spans a spectrum from simple text extraction to sophisticated visual understanding:

Document Parsing Strategies

Text Stream (Legacy) Extract characters via libraries like `pypdf`. Fast and cheap, but destroys tables, columns, and reading order.

Heuristic Parsing Rule-based approaches using whitespace and line detection (e.g., `pdfplumber`). Better table support but brittle on borderless or complex tables.

Layout Models (AI) Neural models like **LayoutLM** or **DocLayout-YOLO** that identify structural elements through bounding boxes. Identifies headers versus body text but requires GPU resources.

Vision-First (VLM) Screenshot pages and process through vision-language models like GPT-4V. ''Human-like'' understanding of layout but slow, expensive, and limited by context windows.

0.3.2 Layout Analysis Models

To solve the PDF problem, modern pipelines employ **layout analysis models** that treat the document page as an image---or a hybrid of image and text box coordinates---to identify visual blocks: headers, paragraphs, tables, and figures.

LayoutLM and Successors.. Microsoft's LayoutLM family combines text understanding with spatial awareness. The model receives both the OCR text and the bounding box coordinates for each text segment, allowing it to learn the relationship between content and position. LayoutLMv3 and subsequent models add visual features from the document image itself, enabling recognition of logos, signatures, and other non-textual elements.

DocLayout-YOLO.. For high-throughput pipelines, object detection architectures like YOLO (''You Only Look Once'') have been adapted for document layout analysis. DocLayout-YOLO processes document images in a single forward pass, detecting and classifying regions as headers, paragraphs, tables, figures, or footnotes. This approach trades some accuracy for significantly faster processing.

Azure Document Intelligence.. Microsoft's cloud service provides pre-trained models for common document types (invoices, receipts, contracts) and allows custom model training for specialized formats. The service returns structured JSON with identified fields, tables, and key-value pairs, along with confidence scores and bounding polygons.

0.3.3 Preserving Structure for Downstream Use

Regardless of the parsing approach, the goal is to preserve sufficient structure for downstream tasks:

Structure Preservation Checklist

- **Reading order**: Reconstruct the logical sequence for multi-column layouts.
- **Hierarchy**: Preserve heading levels, section numbering, and nested lists.
- **Object identifiers**: Retain figure/table numbers and their captions.
- **Page references**: Record page numbers for precise citations.
- **Footnotes/endnotes**: Link footnote markers to their content.
- **Cross-references**: Preserve internal document links where possible.

For legal filings and financial disclosures, page numbers and exhibit references are essential for citation. A contract review system that cannot identify ''Section 4.2(a)'' or ''Exhibit B'' loses the precision that practitioners require.

0.3.4 Chunking Strategies for Retrieval

Once structure is identified, documents must be divided into chunks for embedding and retrieval. Naive approaches split by token count, but structure-aware chunking yields better results:

Structure-Aware Chunking

Semantic boundaries Split at section headings rather than arbitrary token counts.

Table isolation Keep tables as atomic units with their captions.

Contextual overlap Include section headers in each chunk for context.

Metadata preservation Attach page numbers, section paths, and document identifiers to

each chunk.

For financial documents, a common pattern is to extract tables into a separate structured index while chunking the narrative text. This allows the retrieval system to search both modalities and combine results during synthesis.

0.3.5 OCR Quality and Preprocessing

When documents arrive as scanned images rather than digital PDFs, OCR quality becomes the critical bottleneck. Poor OCR produces cascading errors through the entire pipeline.

Resolution Requirements.. Scan quality directly impacts OCR accuracy:

- **300 DPI**: Minimum for standard printed text.
- **600 DPI**: Recommended for fine print, footnotes, and legal documents.
- **Grayscale vs. binary**: Grayscale preserves more information for challenging documents.
- **Color**: Required when color carries semantic meaning (redlines, highlights).

Preprocessing Pipeline.. Before OCR, apply image enhancement:

1. **Deskewing**: Correct rotation from scanner misalignment.
2. **Denoising**: Remove speckles and artifacts.
3. **Contrast enhancement**: Improve readability of faded text.
4. **Binarization**: Convert to black-and-white with adaptive thresholding for text regions.
5. **Border removal**: Eliminate scanner edges and margins.

Multi-Language Documents.. International legal and financial documents may contain multiple languages. OCR engines handle this through:

- Language detection per page or region.
- Multi-language model loading for mixed-language pages.
- Script-specific preprocessing (e.g., right-to-left text handling).

0.3.6 Form and Template Extraction

Structured forms---tax returns, loan applications, regulatory filings---require field-level extraction rather than full-text processing.

Field Detection Approaches..

Template matching Define field locations for known form templates. Fast but brittle to layout variations.

Key-value extraction Use AI to identify label-value pairs without predefined templates. More flexible but requires validation.

Checkbox/radio detection Identify selection states in multi-choice fields.

Table extraction Parse tabular regions within forms.

Named Entity Recognition in Forms.. Once text is extracted from fields, NER identifies:

- Names of individuals and organizations
- Dates and temporal expressions
- Monetary amounts and percentages
- Addresses and contact information
- Account numbers and identifiers

Loan Application Processing

A mortgage application processor might:

1. Classify the document as a specific form type (1003, 1008, etc.).
2. Extract borrower information fields (name, SSN, income).
3. Parse employment history tables.
4. Validate field consistency (do incomes sum correctly?).
5. Flag fields with low confidence for human review.

0.3.7 Complex Layout Challenges

Legal and financial documents often feature complex layouts that defeat simple extraction:

Multi-Column Layouts.. Newspapers, academic journals, and some legal briefs use multiple columns. Extraction must:

- Detect column boundaries.
- Determine reading order (down then across, or across then down).

- Handle text that spans columns (headlines, pull quotes).
- Preserve column-specific context (footnotes belong to their column's text).

Footnotes and Endnotes.. Legal documents heavily use footnotes for citations and qualifications. Extraction should:

- Detect footnote markers in body text.
- Locate corresponding footnote content (often at page bottom or document end).
- Link markers to content for downstream processing.
- Preserve the footnote/body relationship in chunking.

Marginal Annotations.. Contracts and legal filings may include:

- Handwritten margin notes and initials.
- Printed section numbers in margins.
- Change tracking marks (strikethrough, insertions).
- Bates stamps and exhibit markers.

Embedded Objects.. Documents may contain:

- Attached exhibits as embedded PDFs.
- Images and photographs inline with text.
- Signatures and stamps.
- QR codes and barcodes with encoded information.

Complex Layout Processing

- Use layout analysis models to segment before extraction.
- Preserve document hierarchy (sections, subsections, paragraphs).
- Maintain links between footnote markers and content.
- Extract embedded objects to separate processing pipelines.
- Log extraction confidence for quality assurance.

0.4 Tables and Charts

Tables and charts represent high-density information that requires specialized handling. A financial statement table, regulatory schedule, or litigation exhibit contains relationships between cells, headers, and footnotes that simple text extraction destroys. Similarly, charts encode trends and comparisons visually that cannot be reconstructed from OCR alone.

0.4.1 Table Extraction Strategies

Table extraction has evolved from rule-based approaches to sophisticated vision-language methods:

Heuristic Parsers.. Libraries like `pdfplumber` and `Camelot` detect tables through line detection and whitespace analysis. These work well for clean, bordered tables but fail on borderless tables, merged cells, or complex headers common in legal and financial documents.

Vision-Based Extraction.. State-of-the-art approaches utilize vision-language models (VLMs) to parse tables. Instead of reconstructing the table from text coordinates, the system sends an image of the table to a VLM with a prompt to ‘‘transcribe this table to Markdown’’ or ‘‘convert to HTML.’’ This preserves merged cells and complex headers far better than heuristic parsers.

Vision-Based Table Extraction Prompt

```
Extract this table to Markdown format.
Preserve:
- All merged cells (span them appropriately)
- Header hierarchy (multi-row headers)
- Footnote markers (superscript numbers)
- Currency symbols and units
Return only the Markdown table, no explanation.
```

Structured Output for Tables.. When tables must integrate with downstream systems, request structured output:

- **JSON with metadata**: Include row/column headers, units, footnotes, and source page numbers.
- **CSV with context**: Preserve column types and include a header comment with table caption.
- **HTML tables**: Maintain cell spans and styling for complex structures.

0.4.2 Chain-of-Table Reasoning

For reasoning over tables, the **Chain-of-Table** framework dynamically plans operations to navigate a table. Rather than ingesting the whole table into the context, the model iteratively generates operations to create a virtual, simplified table that answers the specific query.

Chain-of-Table Operations

Filter Select rows matching criteria: ‘‘Year > 2020’’

Select Choose specific columns: ‘‘Revenue, Net Income’’

Aggregate Compute sums, averages, or counts

Sort Order by column values

Join Combine with another table on a key column

This mimics how an analyst works with a spreadsheet---progressively narrowing the data to answer a question rather than overwhelming the model with the entire table.

0.4.3 Chart Understanding

Charts are often ignored in text-based retrieval systems, yet they encode critical information in legal and financial documents: stock price trends, market share comparisons, revenue projections.

The CHARGE Framework.. The Chart-based Question Answering Generation (CHARGE) framework extracts keypoints from charts and verifies them against the text to generate QA pairs. This ensures the model can ‘‘read’’ the data trends visually represented in the document, allowing users to ask questions like ‘‘What was the trend in Q3 according to the bar chart?’’

Chart-to-Table Conversion.. One practical approach converts charts to structured data:

1. Send the chart image to a VLM with instructions to extract the underlying data.
2. Request output as a table (CSV or JSON) with axis labels and values.
3. Store both the extracted data and a link to the original chart image.
4. During synthesis, the model can reference either the extracted data or describe the visual.

Chart Extraction Limitations

Be cautious when reconstructing data from charts:

- **Precision loss**: Values read from axis positions are approximate.
- **Missing data points**: Not all values may be visible or labeled.
- **Scale ambiguity**: Logarithmic or truncated axes affect interpretation.
- **Prefer originals**: When available, use the source data files rather than chart extraction.

0.4.4 Multimodal Embeddings for Tables and Charts

Once content is extracted, it must be indexed for retrieval. Two architectural approaches dominate:

Unified Embeddings.. Models like **CLIP** (Contrastive Language-Image Pre-training) and **SigLIP** project images and text into a shared vector space. This allows cross-modal retrieval: a user can type a text query (''Show me the graph of rising interest rates'') and retrieve the relevant image from a slide deck.

Late Fusion.. A robust architecture often employs **late fusion**. Instead of a single embedding space, the system maintains separate indices for text and images (using specialized models for each). During retrieval, candidates are fetched from both indices, and a re-ranking model fuses the scores to present the most relevant mixed-media results.

When to Use Each Approach

Unified embeddings Best for general cross-modal search where text and images are loosely related.

Late fusion Preferred when the nuance of a specific modality (e.g., OCR text inside an image) is critical, or when you need fine-grained control over retrieval weights.

For financial documents with complex tables, late fusion often outperforms unified embeddings because it can leverage specialized table understanding models alongside general-purpose text embedders.

0.5 Audio and Transcripts

Extending retrieval-augmented generation to temporal media (audio and video) introduces the dimension of time. A retrieved result is not just a ''document'' but a specific time span within a media file. For legal and financial practitioners, this means earnings calls, depositions, regulatory hearings, and training recordings become searchable and quotable with timestamp precision.

0.5.1 Audio RAG Pipelines

Audio RAG pipelines depend on the quality of automatic speech recognition (ASR) and the preservation of temporal metadata:

Transcription with Timestamps.. Models like **Whisper** (OpenAI) and **AssemblyAI** convert audio to text while preserving word-level or segment-level timestamps. When a relevant chunk is found during retrieval, the system maps the text back to the original timestamps, allowing the user to ''jump to'' the exact moment in the audio player.

Speaker Diarization.. Crucially, the transcription step must include **speaker diarization**---identifying who is speaking. ‘‘Speaker A said X’’ is semantically different from ‘‘Speaker B said X.’’ In a deposition or earnings call, attributing statements to the correct speaker is essential for accurate analysis.

Audio RAG Pipeline Components

1. **Ingestion**: Audio files processed through ASR with diarization enabled.
2. **Segmentation**: Text chunked by semantic breaks, speaker turns, or silence rather than arbitrary token counts.
3. **Embedding**: Transcript segments embedded with speaker and timestamp metadata.
4. **Retrieval**: Query matches return text plus temporal coordinates.
5. **Synthesis**: Response includes citations with timestamps and optional audio playback links.

Error Rates and Mitigation.. ASR is imperfect. Technical terminology, proper names, and accented speech increase word error rates (WER). For legal and financial applications:

- Provide custom vocabularies (company names, legal terms) to the ASR system.
- Consider human review for high-stakes transcripts (depositions, regulatory testimony).
- Retain the original audio alongside transcripts for verification.
- Display confidence scores where available to flag uncertain passages.

0.5.2 Video Understanding and Retrieval

Video RAG treats video as a sequence of visual frames synchronized with an audio track, enabling queries that span both modalities.

Dual-Stream Indexing.. A comprehensive video RAG system indexes both:

- **Transcript vectors**: What was said (from ASR with diarization).
- **Visual frame descriptions**: What was shown (from keyframe extraction and VLM captioning).

A user query searches both streams, allowing questions like ‘‘Find the scene where the speaker discusses quarterly revenue while showing the bar chart.’’

Keyframe Extraction.. Keyframes are extracted at regular intervals (e.g., 1 frame per second) or at scene changes. Each frame is processed by a VLM to generate textual descriptions (‘‘scene

graphs'') or embedded directly using CLIP. For legal and financial video---training materials, recorded presentations, regulatory hearings---meaningful frames often coincide with slide transitions.

VideoRAG Architecture.. Advanced frameworks like **VideoRAG** employ a dual-channel architecture with ''Graph-based Textual Knowledge Grounding'' to transform visual signals into structured text representations while preserving temporal dependencies. This allows complex queries that span both audio and visual content.

Multimodal Video Query

Query: ''Find where the CFO discusses the accounting change while the slide shows the impact table.''
System behavior:

1. Search transcript for ''accounting change'' + speaker ''CFO''
2. Search visual index for ''table'' or ''impact''
3. Intersect temporal windows to find overlapping segments
4. Return video clips with start/end timestamps

0.5.3 Practical Considerations

Storage and Streaming.. Video and audio files are large. Systems typically:

- Store original media in object storage (S3, Azure Blob).
- Generate and index transcripts/descriptions separately.
- Stream relevant segments via FFMPEG or cloud media services.
- Return playback links with timestamp parameters rather than downloading entire files.

Privacy and Access Control.. Audio and video often contain sensitive content---voices are biometric identifiers, and recordings may capture privileged communications. Apply the privacy controls discussed in Section 0.6 before ingestion:

- Redact or exclude segments containing privileged discussions.
- Apply speaker-level access controls where content is speaker-specific.
- Consider whether transcripts alone (without audio) suffice for the use case.

Audio/Video RAG Best Practices

- Always preserve timestamp-to-text mappings for citation.
- Enable speaker diarization for multi-party recordings.
- Provide custom vocabularies for domain-specific terminology.
- Retain original media for verification of AI-generated transcripts.
- Apply access controls at the segment level where sensitivity varies.

0.5.4 ASR Error Handling and Quality Control

Automatic speech recognition is imperfect, and legal and financial applications require explicit quality management.

Word Error Rate Estimation.. WER measures transcription quality but requires ground truth for calculation. In production:

- Use confidence scores from the ASR engine as proxies.
- Flag low-confidence segments for review.
- Sample and manually review a percentage of transcripts to estimate system-wide WER.
- Track WER by speaker, audio quality, and domain to identify systematic issues.

Confidence Thresholds.. Configure thresholds to balance automation and quality:

- **High confidence (>0.9)**: Accept without review for routine use.
- **Medium confidence (0.7--0.9)**: Flag for optional review; acceptable for search but not citation.
- **Low confidence (<0.7)**: Require human review before reliance; mark uncertain passages visually.

Domain-Specific Vocabulary.. Legal and financial terminology often confuses general ASR models:

- **Custom vocabulary lists**: Company names, product names, legal terms, ticker symbols.
- **Pronunciation hints**: Guide recognition for unusual names or acronyms.
- **Boosted phrases**: Increase likelihood of domain-specific terms appearing.
- **Post-processing correction**: Rule-based substitution for common errors.

Financial Earnings Call Vocabulary

For an earnings call transcription system:

- Boost company name, subsidiary names, and executive names.
- Include accounting terms: ‘‘EBITDA,’’ ‘‘goodwill impairment,’’ ‘‘diluted EPS.’’
- Add industry-specific terminology: ‘‘basis points,’’ ‘‘comps,’’ ‘‘guidance.’’
- Handle ticker symbols and numeric expressions.

0.5.5 Video Frame Analysis

Beyond the audio track, video content carries visual information that may be critical for understanding.

Keyframe Selection Strategies.. Not every frame needs processing. Selection strategies include:

- **Fixed interval**: Sample every N seconds (simple but may miss important moments).
- **Scene change detection**: Identify visual transitions and sample at boundaries.
- **Motion analysis**: Capture frames when significant visual change occurs.
- **Audio-aligned**: Select frames when speakers change or key topics are mentioned.

Frame Understanding.. Process selected keyframes through vision-language models to extract:

- Scene descriptions (‘‘Speaker at podium with presentation slide’’).
- Text visible in frame (slide content, whiteboard text, on-screen graphics).
- Object detection (people, documents, equipment).
- Face identification (if authorized and relevant).

Exhibit Detection in Depositions.. Video depositions often include exhibit presentations. Detection should:

- Identify when documents are shown on camera.
- Capture exhibit images for separate processing.
- Link exhibit appearances to transcript timestamps.
- Note when witnesses are reviewing documents versus speaking.

0.5.6 Slide Deck Extraction

Recorded presentations often feature slide decks that carry structured information separate from the spoken narrative.

Slide Detection and Extraction..

- Detect slide transitions in the video stream.
- Capture representative frame for each slide.
- OCR slide text content.
- Extract charts and diagrams for visual analysis.

Slide-Transcript Synchronization.. Align extracted slides with the spoken transcript:

- Map slide transitions to transcript timestamps.
- Associate spoken content with the visible slide.
- Enable queries like ‘‘What did the speaker say about slide 5?’’

Speaker Notes and Metadata.. If the original presentation file is available (PowerPoint, Google Slides):

- Extract speaker notes as additional context.
- Preserve slide titles and section structure.
- Capture embedded links and references.
- Index slide-level metadata alongside video segments.

0.5.7 Real-Time vs. Batch Processing

Audio and video processing can occur live or after recording, with different trade-offs.

Real-Time Transcription.. Live transcription for hearings, meetings, or trading floors:

- Lower latency requirements (seconds, not minutes).
- Streaming ASR models that process audio chunks incrementally.
- Immediate display for participants (accessibility, record-keeping).
- Reduced accuracy compared to batch processing (less context available).

Batch Processing for Discovery.. Historical recordings processed in bulk:

- Higher accuracy through multiple passes and post-processing.
- Full audio context available for each segment.
- Batch speaker diarization with cross-recording speaker linking.
- Parallelization across compute resources.

Latency-Accuracy Trade-offs.. The fundamental trade-off:

- Real-time: Lower latency, lower accuracy, immediate availability.
- Batch: Higher latency, higher accuracy, suitable for archival and search.
- Hybrid: Real-time draft followed by batch refinement for permanent record.

Processing Mode Selection

Real-time Accessibility for live events, trading floor monitoring, meeting assistance.

Batch E-discovery, historical research, compliance review, permanent transcripts.

Hybrid Live captioning refined overnight for searchable archives.

0.6 Privacy and Redaction

Data leakage is a primary concern in multimodal RAG systems, while misinformation is a concern in generation. Before data enters the vector database or the LLM context window, it must be scrubbed of personally identifiable information (PII), privileged content, and other sensitive material. Equally important, the provenance of AI-generated content must be trackable.

0.6.1 PII Detection and Redaction

The Challenge.. Documents processed through multimodal pipelines often contain:

- **Direct identifiers**: Names, Social Security numbers, account numbers.
- **Quasi-identifiers**: Dates, locations, and demographic details that enable re-identification.
- **Sensitive categories**: Health information, financial data, legal case details.
- **Embedded PII**: Information within images, scanned forms, or handwritten notes.

Microsoft Presidio.. Presidio is an open-source framework for detecting, redacting, masking, and anonymizing sensitive data. It combines:

- **Pattern matching**: Regular expressions for structured identifiers (SSN, phone numbers, credit cards).
- **Named entity recognition (NER)**: Machine learning models to identify names, organizations, and locations.
- **Configurable anonymizers**: Replace, mask, hash, or encrypt detected entities.
- **Extensibility**: Custom recognizers for domain-specific identifiers (case numbers, account formats).

Presidio Redaction Pipeline

A document processing pipeline might:

1. Extract text from PDF using layout analysis.
2. Pass text through Presidio's analyzer to detect PII entities.
3. Apply anonymizers: replace names with tokens, mask account numbers.
4. Embed and index the redacted text.
5. Store mapping between tokens and original values in a secure vault (if reversible redaction is needed).

Image-Based PII.. For scanned documents and images:

- Run OCR to extract text, then apply text-based PII detection.
- Use bounding box coordinates to redact regions in the original image.
- Consider visual PII (faces, signatures) that text-based methods miss.
- Presidio Image Redactor extends the framework to handle images directly.

0.6.2 Privilege and Confidentiality

Beyond PII, legal and financial workflows must protect:

- **Attorney-client privilege**: Communications protected from disclosure.
- **Work product doctrine**: Attorney mental impressions and legal strategy.
- **Trade secrets**: Proprietary business information.
- **Material non-public information (MNPI)**: Information that could affect securities prices.

Privilege in AI Systems

Including privileged content in a shared vector database or sending it to an external LLM API may waive privilege. Design systems to:

- Segregate privileged content into separate indices with strict access controls.
- Use on-premises or private cloud deployments for sensitive processing.
- Implement privilege review workflows before ingestion.
- Log all access to privileged content for audit purposes.

0.6.3 Content Authenticity and Provenance

As AI generates increasingly realistic content, tracking provenance becomes critical. The **Content Authenticity Initiative (CAI)** and **C2PA** (Coalition for Content Provenance and Authenticity) standards address this need.

Content Credentials.. C2PA enables cryptographic signing of media files. This metadata (''Content Credentials'') travels with the file, proving:

- **Origin**: Whether content was AI-generated, camera-captured, or edited.
- **Editing history**: What modifications were applied and by whom.
- **Tool chain**: Which software or AI models were involved.

This provides a ''digital nutrition label'' that allows consumers to verify the provenance of the content they are viewing.

Application to Legal and Finance.. Content Credentials are particularly relevant for:

- **Evidence authenticity**: Establishing the chain of custody for digital evidence.
- **AI-generated disclosures**: Marking synthetic content in regulatory filings.
- **Document integrity**: Proving that a contract or filing has not been tampered with.
- **Audit trails**: Demonstrating the provenance of AI-assisted analysis.

0.6.4 Redaction Governance

Effective redaction requires governance beyond the technical implementation:

Redaction Governance Checklist

- **Policy documentation**: Define what must be redacted and under what circumstances.
- **Version control**: Track changes to redaction rules over time.
- **Exception handling**: Document when and why redaction was overridden.
- **Audit logging**: Record who performed redactions, when, and what was affected.
- **Reversibility decisions**: Determine if redaction should be reversible and secure the mapping.
- **Quality assurance**: Sample and review redacted output for completeness.

Integration with Evidence Records.. Redaction events should be captured in the canonical evidence record (see Chapter 3). When a document is processed:

- Log the redaction rules applied (version, configuration).
- Record entities detected and actions taken.
- Preserve checksums of both original and redacted content.
- Link to the redaction policy governing the action.

This ensures that any downstream analysis can be traced back to the original data with full understanding of what was removed and why.

0.6.5 Visual Privacy Patterns

Multimodal content introduces privacy risks that text-based detection misses entirely.

Face Detection and Blurring.. Images and video containing faces require special handling:

- **Detection models**: Use face detection (not recognition) to identify face regions.
- **Blurring techniques**: Gaussian blur, pixelation, or solid overlay for varying anonymization strength.
- **Temporal consistency**: In video, track faces across frames to maintain consistent blurring.
- **Edge cases**: Profile views, partial occlusion, and small faces require robust models.

Biometric Data Regulations.. Face images constitute biometric data under several regulatory frameworks:

- **Illinois BIPA**: Requires informed consent before collecting biometric identifiers.

- **GDPR Article 9**: Classifies biometric data as special category requiring explicit consent.
- **CCPA/CPRA**: Includes biometric information in the definition of sensitive personal information.
- **Sector-specific rules**: Financial and healthcare regulations may impose additional requirements.

Visual Signature and Handwriting.. Handwritten content carries identifying characteristics:

- Signatures are personally identifying and may constitute biometric data.
- Handwriting style can enable re-identification even when names are redacted.
- Consider redacting handwritten annotations in shared datasets.
- Document whether handwriting analysis was performed and by what method.

Visual PII in Legal Discovery

During e-discovery, photographs and video may contain faces of witnesses, victims, or bystanders. Before sharing with opposing counsel or including in public filings:

- Identify all individuals whose faces appear.
- Determine if protective orders require redaction.
- Apply consistent blurring across all appearances.
- Document the redaction in the privilege log or production notes.

0.6.6 MNPI Detection Workflows

Material Non-Public Information requires specialized detection beyond general PII frameworks.

The MNPI Challenge.. Unlike PII with structural patterns (SSN formats, phone numbers), MNPI is contextual:

- **Materiality**: Information that a reasonable investor would consider important.
- **Non-public**: Not yet disclosed through official channels.
- **Context-dependent**: The same statement may be MNPI before an announcement and public after.
- **Temporal sensitivity**: MNPI status changes when information is disclosed.

Detection Approaches.. Automated MNPI detection combines multiple signals:

- **Entity recognition**: Identify company names, ticker symbols, and executive names.

- **Event classification**: Detect references to earnings, M&A, product launches, regulatory actions.
- **Temporal analysis**: Cross-reference against public disclosure calendars.
- **Source classification**: Distinguish between public sources and internal communications.

Integration with Compliance Workflows.. MNPI detection integrates with broader compliance systems:

- **Information barriers**: Enforce separation between deal teams and trading desks.
- **Watch lists and restricted lists**: Flag securities subject to trading restrictions.
- **Attestation workflows**: Require certification before accessing sensitive content.
- **Audit trails**: Log all access to potentially MNPI-containing documents.

MNPI in Earnings Call Processing

When processing earnings call recordings:

1. Check if the call is from a public or private company.
2. Verify the call has been publicly webcast or transcripted.
3. For private companies, classify content as potentially MNPI.
4. Apply access controls based on user's compliance status.
5. Log all queries and retrieved content for compliance review.

0.6.7 Cross-Border Data Considerations

Multimodal processing often involves cross-border data transfers that trigger additional regulatory requirements.

Data Localization Requirements.. Various jurisdictions impose data residency requirements:

- **GDPR**: Restricts transfers outside the EU/EEA without adequacy decisions or safeguards.
- **China PIPL**: Requires security assessments for certain cross-border transfers.
- **Russia**: Mandates storage of Russian citizens' data within Russia.
- **Sector-specific rules**: Financial regulators may impose additional localization requirements.

Transfer Mechanisms.. When processing must occur across borders:

- **Standard Contractual Clauses**: GDPR-approved contract terms for international transfers.

- **Binding Corporate Rules**: Internal policies for multinational organizations.
- **Consent**: Explicit consent from data subjects (limited applicability).
- **Necessity exceptions**: Transfers necessary for legal claims or vital interests.

Cloud Provider Selection.. Choice of cloud infrastructure affects compliance:

- **Regional deployment**: Process data in the region where it originates.
- **Provider certifications**: Verify SOC 2, ISO 27001, and sector-specific certifications.
- **Sub-processor management**: Understand where provider's sub-contractors process data.
- **Government access**: Consider jurisdiction of cloud provider for government access requests.

Cross-Border Checklist for Multimodal AI

- Map data flows: Where does content originate, process, and store?
- Identify applicable laws: Which jurisdictions' privacy laws apply?
- Implement transfer mechanisms: SCCs, BCRs, or other safeguards.
- Configure regional processing: Deploy models and storage in appropriate regions.
- Document compliance: Maintain records of assessments and safeguards.
- Monitor for changes: Regulatory landscape evolves rapidly.

0.7 Synthesis

Multimodal RAG represents the maturation of retrieval-augmented generation from a text-processing technique to a comprehensive perception system. By addressing document structure, tables and charts, images, audio, video, and privacy safeguards, you can build workflows that match the multimodal reality of legal and financial practice.

0.7.1 Core Technical Themes

This chapter has developed several interconnected themes:

Structure Preservation.. The PDF problem---extracting meaning from documents designed for visual rendering---requires moving beyond naive text extraction. Layout analysis models (LayoutLM, DocLayout-YOLO), structure-aware chunking, and metadata preservation ensure that downstream systems understand not just what text says, but how it relates to tables, headers, and cross-references.

Multimodal Integration.. Legal and financial content spans modalities: scanned documents require OCR, charts require visual understanding, depositions require audio transcription with speaker diarization, and video presentations require synchronized analysis of slides and speech. Unified embeddings (CLIP, SigLIP) or late fusion architectures enable queries that span these modalities.

Privacy by Design.. PII detection, privilege protection, and MNPI handling must be integrated from ingestion, not bolted on afterward. Visual privacy (faces, signatures, handwriting) requires specialized detection. Cross-border data flows add complexity that must be addressed in pipeline architecture.

Provenance and Authenticity.. As AI generates content indistinguishable from human-created material, content credentials (C2PA) and evidence records become essential. Every transformation should be logged, enabling downstream verification and audit.

0.7.2 Key Takeaways

Chapter Takeaways

1. **Tables require special handling**: Naive text extraction destroys tabular structure. Use table-specific extraction, serialization (Markdown, HTML), and reasoning techniques.
2. **Audio and video add temporal dimensions**: Retrieval returns time spans, not documents. Preserve timestamp-to-text mappings for citation.
3. **OCR quality cascades through pipelines**: Low-resolution scans, poor preprocessing, or missing domain vocabularies create errors that propagate to embedding and retrieval.
4. **Privacy is multimodal**: Text-based PII detection misses faces, signatures, and contextual MNPI. Layer visual and contextual privacy analysis.
5. **Structure-aware chunking outperforms token counting**: Split at semantic boundaries, keep tables atomic, include headers for context.

0.7.3 Integration Patterns

The components discussed in this chapter work together in layered pipelines:

1. **Ingestion layer**: Documents enter the system and are classified by type (PDF, image, audio, video).
2. **Preprocessing layer**: Layout analysis, table extraction, OCR, and ASR transform raw content into structured text with metadata.
3. **Privacy layer**: PII detection and redaction sanitize content before it enters shared indices or

external APIs.

4. **Embedding layer**: Content is vectorized---potentially through multiple specialized embedders (text, image, table).
5. **Indexing layer**: Vectors and metadata are stored with provenance information.
6. **Retrieval layer**: Queries search across modalities, with late fusion combining results.
7. **Synthesis layer**: Retrieved content is presented to the LLM with appropriate context for generation.

0.7.4 Architectural Decisions

When designing multimodal pipelines, key architectural decisions include:

- **Parsing strategy**: Heuristic, AI-based layout models, or vision-first (VLM)?
- **Embedding architecture**: Unified multimodal embeddings or late fusion?
- **Privacy approach**: Pre-ingestion redaction, access controls, or both?
- **Media handling**: Stream from source or cache processed segments?
- **Provenance depth**: Minimal logging or full W3C PROV-O lineage?
- **Real-time vs. batch**: Lower latency with reduced accuracy, or higher accuracy with batch processing?

The right answers depend on your accuracy requirements, latency constraints, cost sensitivity, and regulatory obligations.

0.7.5 What This Chapter Did Not Cover

Several related topics fall outside this chapter's scope:

- **Fine-tuning multimodal models**: Training custom VLMs or document understanding models for specialized domains.
- **Real-time streaming architectures**: Processing live video feeds or continuous audio streams at scale.
- **3D and spatial content**: CAD files, BIM models, and other spatial data formats.
- **Specific tool implementations**: Detailed configuration of Tesseract, Azure AI, or specific ASR systems.
- **Litigation hold and preservation**: E-discovery-specific workflows for document collection and preservation.

0.7.6 Connecting to Other Chapters

Multimodal Perception in Agentic Systems

Document processing and multimodal understanding become *perception capabilities* in agentic systems. Chapter 7's Perception question addresses how agents access and interpret diverse information sources, while the Governance question covers privacy controls and data isolation requirements. The pipelines described here form the sensory apparatus through which agents perceive their documentary environment.

Chapter 3: Retrieval-Augmented Generation.. The multimodal techniques here extend the text-based RAG foundations from Chapter 3. Late fusion, multimodal embeddings, and temporal retrieval build on vector search and retrieval evaluation concepts.

Chapter 5: Prompt Design and Evaluation.. Prompts that incorporate multimodal context---table representations, image descriptions, transcript excerpts---require the prompt engineering techniques covered in Chapter 5.

Chapter 6: Agentic Systems.. Agents use multimodal perception to understand their environment. Document analysis, chart interpretation, and audio transcription become perception tools that agents invoke during task execution.

0.7.7 Looking Forward

With multimodal ingestion in place, the next challenge is designing prompts and evaluation frameworks that leverage these capabilities effectively. Chapter 5 treats prompt design, strategy selection, evaluation, and optimization as an engineering discipline---applying structured thinking to the interface between human intent and model behavior.

0.8 Further Learning

This section provides an annotated guide to primary sources and resources for readers who wish to deepen their understanding of multimodal document processing, retrieval, and privacy. We organize resources by topic, with brief annotations explaining the relevance and accessibility of each source.

0.8.1 Document Layout and Structure

LayoutLM Family.. The LayoutLM series from Microsoft Research represents the state of the art in document understanding:

- **LayoutLM** (Xu et al., 2020): Introduced the combination of text and layout information in a

pre-trained transformer. Foundational for understanding modern document AI.

- **LayoutLMv2** (Xu et al., 2021): Added visual features from document images, enabling recognition of non-textual elements.
- **LayoutLMv3** (Huang et al., 2022): Unified text-image pre-training with improved performance across document understanding tasks.

Object Detection Approaches.. For high-throughput document processing:

- **DocLayout-YOLO**: Adapts YOLO object detection for document layout analysis. Useful for fast identification of headers, paragraphs, tables, and figures.
- **PaddleOCR**: Baidu's open-source OCR system with layout analysis capabilities.
- **Tesseract LSTM**: The open-source standard, with neural network-based recognition.

Commercial Solutions.. Enterprise-grade document processing:

- **Azure AI Document Intelligence**: Microsoft's cloud service with pre-trained models for invoices, receipts, and custom documents.
- **Amazon Textract**: AWS service for OCR, table extraction, and form processing.
- **Google Document AI**: Cloud-based document understanding with specialized processors.

0.8.2 Table and Chart Understanding

Table Reasoning.. Moving beyond table extraction to table understanding:

- **Chain-of-Table** (Wang et al., ICLR 2024): Framework for table reasoning through iterative operations. Demonstrates that step-by-step reasoning outperforms single-pass table ingestion.
- **Table-GPT**: Research on specialized table understanding models and prompting strategies.
- **TAPAS/TAPEX**: Google's table pre-training approaches for question answering over tables.

Chart Understanding.. Extracting information from data visualizations:

- **CHARGE**: Chart-based Question Answering Generation framework for extracting and verifying information from visualizations.
- **ChartQA**: Benchmark for visual question answering on charts.
- **MatCha/DePlot**: Google's chart understanding models that convert charts to tables.

0.8.3 Audio and Video Processing

Speech Recognition.. Automatic speech recognition for transcription:

- **Whisper** (OpenAI): Open-source ASR supporting 99 languages with word-level timestamps. The practical standard for most applications.
- **AssemblyAI**: Commercial ASR with speaker diarization and real-time streaming.
- **Deepgram**: Enterprise ASR with customization for domain vocabulary.
- **AWS Transcribe**: Amazon's ASR service with custom vocabulary support.

Speaker Diarization.. Identifying who is speaking:

- **pyannote.audio**: Open-source speaker diarization toolkit in Python.
- **NeMo**: NVIDIA's toolkit includes speaker diarization models.
- **Commercial APIs**: AssemblyAI, Rev.ai, and others provide diarization as a service.

Video Understanding.. Processing video content:

- **VideoRAG**: Frameworks for long-context video understanding with dual-channel retrieval.
- **Twelve Labs**: Commercial video understanding and search API.
- **Video-LLaVA**: Open-source video understanding models.

0.8.4 Multimodal Embeddings

Vision-Language Models.. Unified understanding of images and text:

- **CLIP** (OpenAI): Contrastive Language-Image Pre-training. Foundational model for cross-modal retrieval.
- **SigLIP** (Google): Sigmoid Loss for Language Image Pre-Training. Improved calibration for retrieval tasks.
- **OpenCLIP**: Open-source CLIP implementations with various model sizes.

Multimodal RAG Architectures.. Combining retrieval across modalities:

- **Late Fusion**: Research comparing unified embeddings versus modality-specific indices with score fusion.
- **ColPali/ColQwen2**: Dense retrieval models for document understanding.
- **BLIP-2**: Bootstrapped vision-language pre-training for multimodal understanding.

0.8.5 Privacy and Content Authenticity

PII Detection and Redaction..

- **Microsoft Presidio**: Open-source PII detection and anonymization framework. Supports text and images, extensible for custom entity types.
- **spaCy NER**: Named entity recognition for identifying names, organizations, and locations.
- **AWS Comprehend**: Amazon's NLP service with PII detection capabilities.

Content Authenticity.. Provenance and tamper-evidence:

- **C2PA Specification**: The Coalition for Content Provenance and Authenticity technical specification for cryptographic content credentials.
- **Content Authenticity Initiative**: Adobe-led consortium developing tools and standards for content provenance.
- **c2pa-rs**: Open-source Rust implementation of the C2PA specification.

0.8.6 Practical Guides and Documentation

Beyond research papers, practitioners benefit from implementation guides:

- **Unstructured.io**: Open-source library for document parsing with practical documentation. `https://unstructured.io`
- **LangChain Document Loaders**: Framework components for loading and parsing diverse document types. `https://docs.langchain.com`
- **LlamaIndex**: Data connectors and document processing for RAG applications. `https://docs.llamaindex.ai`
- **Azure AI Documentation**: Comprehensive guides for Document Intelligence, including table extraction and custom models. `https://learn.microsoft.com/azure/ai-services/document-intelligence`

0.8.7 Common Misconceptions

Before concluding, several common misconceptions warrant clarification:

Misconception: VLMs eliminate the need for specialized extraction.. Vision-language models can understand documents directly, but specialized extraction pipelines remain more reliable, faster, and cheaper for production workloads. VLMs complement rather than replace structured extraction.

Misconception: OCR is a solved problem.. Modern OCR achieves high accuracy on clean printed text, but handwriting, low-quality scans, and complex layouts still challenge even the best systems. Quality varies significantly by document type.

Misconception: Audio transcription is accurate enough for legal use.. ASR error rates, while low for clear speech, can exceed 10--20% for technical terminology, accented speech, or poor audio quality. Legal and financial applications require human review for high-stakes content.

Misconception: Redaction is reversible if you keep the mapping.. While technical reversibility is possible, legal and regulatory requirements may prohibit retaining mappings. Design for true deletion when required.

0.8.8 Exercises for Practitioners

To solidify understanding of the concepts in this chapter, we recommend the following exercises:

Exercise 1: Table Extraction Comparison.. Take a complex financial table (earnings report, balance sheet) and process it through multiple extraction methods:

1. Extract using basic text extraction (pypdf).
2. Extract using layout analysis (Unstructured, Azure).
3. Extract using VLM description (GPT-4V, Claude).
4. Compare outputs: Which preserves structure best? Which handles merged cells?

Exercise 2: ASR Quality Assessment.. Record a simulated earnings call with technical terminology:

1. Transcribe using Whisper or another ASR system.
2. Create a ground truth transcript manually.
3. Calculate WER and identify systematic errors.
4. Add custom vocabulary and measure improvement.

Exercise 3: Privacy Pipeline Design.. Design a document processing pipeline for a hypothetical law firm:

1. Document types to process (contracts, correspondence, pleadings).
2. PII categories to detect and handling for each.
3. Privilege markers to identify and routing rules.

4. Access control model for processed content.

Exercise 4: Multimodal Query Design.. For a video deposition with exhibits:

1. Design the indexing strategy (transcript, exhibits, video frames).
2. Write 5 queries that span multiple modalities.
3. Describe how the system would retrieve and present results.
4. Identify what metadata is needed for accurate citation.

Conclusion

The transition from text-only to multimodal RAG represents a qualitative shift in what AI systems can perceive and process. By applying layout analysis to documents, specialized extraction to tables and charts, ASR with diarization to audio and video, and rigorous privacy controls throughout, you can build systems that match the multimodal reality of legal and financial practice.

Key takeaways from this chapter:

- **Structure matters**: The ‘‘PDF problem’’ destroys semantic relationships. Layout analysis models preserve the structure that practitioners need for accurate citation and analysis.
- **Tables require special handling**: Whether through heuristic parsers, vision-based extraction, or Chain-of-Table reasoning, tables must be treated as first-class objects, not flattened text.
- **Time is a dimension**: Audio and video RAG adds temporal coordinates to retrieval, enabling precise citation of spoken content.
- **Privacy is non-negotiable**: PII detection, redaction, and privilege protection must occur before content enters shared systems or external APIs.
- **Provenance enables trust**: Content Credentials and evidence records establish the chain of custody for AI-processed content.

With multimodal ingestion capabilities in place, the next challenge is the human-AI interface: how do we communicate intent to these systems effectively? Chapter 5 addresses prompt design, evaluation, and optimization as engineering disciplines---applying the same rigor to the interface layer that we have applied to structured outputs, tool use, and multimodal processing.