MULTIMODAL AI MASTERY

Vision, Voice, Video & Documents as Strategic Business Context

INSIDE THIS GUIDE:

- → GPT-5, Claude 4 & Gemini 2.5 compared
- → Image prompting frameworks
- → Video-as-Code prototype method
- → The Multimodal Context Pyramid
- → Audio & meeting intelligence
- → Failure modes & fixes for each modality

3

The Text-Only Era Is Over

In 2026, the most capable AI models process text, images, audio, video, and documents simultaneously. This is not a gimmick — it is a fundamental shift in what constitutes a "prompt." The professionals who learn to provide rich, multimodal context will produce outputs that text-only users cannot match.

> **THE TMINUSAI AXIOM**
>
> "Context is no longer just words. It is what the AI can see, hear, and read in your documents. Feed the model the same context a human expert would need."

The Three Leading Multimodal Models (March 2026)

Model	Text	Image	Audio	Video	Docs	Strength
GPT-5.2	★★★ ★★	★★★ ★■	★★★ ★■	★★★ ★■	★★★ ★★	Coding, structured analysis, spreadsheets
Claude 4 Sonnet	★★★ ★★	★★★ ★★	★★★ ■■	★★★ ■■	★★★ ★★	Document analysis, nuanced writing, safety
Gemini 2.5 Pro	★★★ ★■	★★★ ★★	★★★ ★★	★★★ ★★	★★★ ★■	Video, audio, long-context multimodal

Note: Star ratings reflect relative strengths as of March 2026. All three models are highly capable across all modalities. Choose based on your primary use case.

SECTION 1 — The Multimodal Context Pyramid

How to Structure Non-Text Inputs

Raw multimodal input without structure produces raw output. The TminusAI Multimodal Context Pyramid (MCP) is a layered framework for anchoring visual and auditory data with explicit intent at every level.

Layer	Input Type	Strategic Use	TminusAI Best Practice
Peak: Temporal Context	Video sequences, screen recordings	UX flows, product demos, training videos	Provide key timestamps. Segment long videos before submitting.
Layer 3: Auditory Nuance	Meeting recordings, voice memos	Decision extraction, sentiment analysis	Provide a "Speaker Key." Pre-transcribe with Whisper for clarity.
Layer 2: Visual Evidence	Screenshots, diagrams, charts, UI	UX audit, data analysis, design review	Use annotated screenshots. Ask model to "OCR all text first."
Base: Textual Anchor	Core instruction and goal	All tasks require this foundation	Always state "Why" and "Goal" before attaching non-text input.

SECTION 2 — Image Prompting Frameworks

Category A: Image Intelligence

Framework 1 — The UI/UX Auditor

Input: A high-resolution screenshot of your product's landing page or dashboard.

```
"Analyze the visual hierarchy of this screenshot. Identify 3 specific friction points that prevent users from clicking the primary CTA. Reference exact UI elements by location. Compare against the 3 principles of high-converting SaaS pages. Provide your audit as: [Observation] -> [Principle Violated] -> [Specific Fix]."
```

Framework 2 — The Data Chart Interpreter

Input: A screenshot of a chart, graph, or dashboard.

```
"You are a data analyst. OCR all text and numbers in this chart first. Then: 1) State the key trend in one sentence. 2) Identify any anomalies or outliers. 3) Suggest the 2 most likely causal explanations. 4) Recommend the next analytical step. Do not speculate beyond what the data shows."
```

Framework 3 — The Document Extractor

Input: A scanned PDF, invoice, contract, or business card.

```
"Extract all structured information from this document. Output as JSON with these keys: [list your fields]. If a field is unclear or absent, mark it null — do not guess. Then summarize the document's primary purpose in one sentence."
```

> **FAILURE MODE: VISUAL HALLUCINATION**
>
> Models can "see" buttons, text, or UI elements that do not exist in low-resolution or cluttered images. Fix: Always request "OCR all text verbatim first" as Step 0 before any analysis. This forces the model to ground its observations in confirmed text.

SECTION 3 — Audio Intelligence

Category B: Audio & Meeting Intelligence

Models like Gemini 2.5 Pro and GPT-5.2 can process audio files directly. For best results with complex multi-speaker recordings, pre-transcribe with Whisper v3 and feed the labeled transcript.

Framework 4 — The Decision Extraction Protocol

Input: A 30-minute recorded meeting (audio file or labeled transcript).

```
"Extract ONLY the following from this meeting: 1) Decisions made (state exactly what was decided and by whom). 2) Action items assigned (owner + deadline). 3) Open questions that were NOT resolved. 4) Conduct a Sentiment Audit: flag any moments of unaddressed tension or disagreement that may resurface. Ignore small talk entirely. Output in structured Markdown."
```

Framework 5 — The Voice Memo Intelligence Digest

Input: Your own voice memo (recorded on phone while commuting or walking).

```
"This is a voice memo I recorded while thinking out loud. Ignore filler words and self-corrections. Extract: 1) The core idea I was developing. 2) Any action items I mentioned. 3) Questions I raised but did not answer. Format as a structured note I can paste into Notion."
```

> **FAILURE MODE: AUDITORY MISINTERPRETATION**
>
> Models miss sarcasm, subtle tone changes, and implied disagreement without explicit instruction. Fix: Add "flag any statements that contradict earlier statements from the same speaker" to any meeting analysis prompt. This forces the model to detect internal inconsistencies.

SECTION 4 — Video Intelligence

Category C: Video as Code & Context

Video analysis is where Gemini 2.5 Pro currently leads. GPT-5.2 also has strong video capabilities. The key is segmentation — never feed an unstructured long video to a model without a timestamp map.

Framework 6 — The Video-as-Code Prototype

Input: A 2-5 minute screen recording of an app or web flow.

```
"Analyze this screen recording. Step 1: Identify the core user flow (what task is the user trying to complete?). Step 2: List every friction point where the user hesitated, re-clicked, or went back. Step 3: Write the React/Tailwind code for a redesigned prototype that eliminates the top 2 friction points. Include inline comments explaining each design decision."
```

Framework 7 — The Training Video Converter

Input: A recorded onboarding or training video.

```
"Convert this training video into a structured written guide. Include: 1) A numbered step-by-step walkthrough. 2) Screenshots descriptions at key steps (describe what should be visible). 3) A FAQ section with the 5 most likely questions a new user would have. 4) A one-page summary I can distribute as a handout."
```

> **FAILURE MODE: CONTEXT OVERLOAD**
>
> Providing too many multimodal inputs simultaneously (e.g., 5 images + 2 PDFs + an audio file) causes "Attention Dilution" — the model under-processes the most critical evidence. Fix: Process one modality at a time. Feed outputs as text context for the next step.

QUICK REFERENCE — Multimodal Cheat Sheet

Multimodal Input Quick Reference

Use Case	Input Type	Best Model	Key Prompt Instruction
Landing page UX audit	Screenshot	GPT-5.2 / Claude 4	"OCR all text first. Identify 3 friction points."
Contract review	PDF document	Claude 4 Sonnet	"Extract all obligations, deadlines, and penalty clauses."
Meeting notes	Audio / transcript	Gemini 2.5 Pro	"Extract decisions, action items, unresolved questions."
UX prototype from recording	Screen recording	Gemini 2.5 Pro / GPT-5.2	"Identify flow. Write React prototype for top 2 fixes."
Invoice/form data extraction	Scanned PDF / image	GPT-5.2 / Claude 4	"Output as JSON. Mark unclear fields as null."
Chart interpretation	Screenshot of graph	GPT-5.2	"OCR numbers first. State key trend. Flag anomalies."
Training video to guide	Video file	Gemini 2.5 Pro	"Convert to numbered walkthrough with step descriptions."

NEXT STEPS

You now have the multimodal toolkit. Continue with Guide 4: The Prompt Scorecard to build a rigorous QA loop that separates professional-grade AI outputs from amateur ones. Available at tminusai.com.