Multimodal AI

When AI Has Eyes and Ears

Analyzing and Creating with Images, Audio, and Video

UBUS 670 | AI for Business Leaders

Day 4 • Week 2 • Monday, March 16, 2026 • 4 hours • Google Gemini

Today's Learning Objectives

By the end of today, you will be able to:

1. Explain how multimodal AI processes images, audio, and video — not just text
2. Analyze business content across multiple modalities (images, audio, video) using Gemini
3. Generate marketing visuals using AI image generation with effective prompts
4. Design a structured multimodal workflow that combines analysis and generation for a business deliverable

Today's Skill:

Multimodal AI

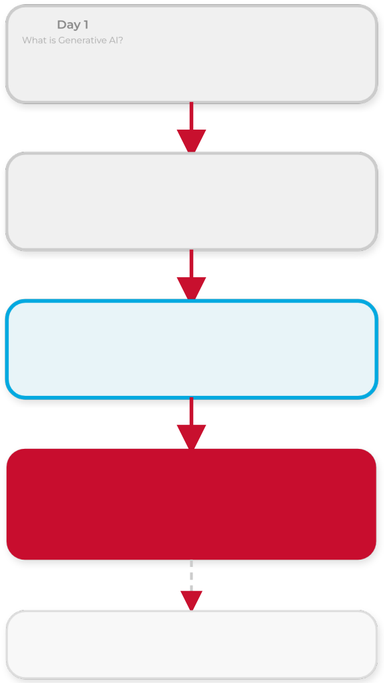
From text → images, audio, video

Quick Recap: Day 3 → Day 4

Day 3: Context Engineering — key takeaways:

- Context engineering builds information environments for AI
- Structure helps — Markdown, JSON, clear formatting improve results
- RAG + Embeddings power semantic search over your own data
- Gemini Gems give you no-code RAG with persistent context
- RAG beats fine-tuning for most business cases

> Bridge: Day 3 taught you to feed AI the right text context — documents, structured formats, Gemini Gems. Today: the context isn't just text anymore. Images, audio, and video are all context. And AI can now create visual content too.



Section 1
Beyond Text

What is Multimodal AI & Why Does It Matter?

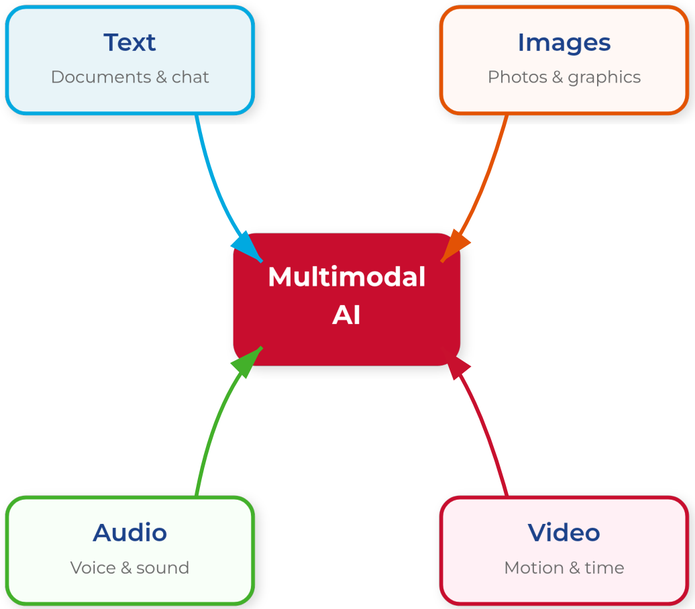
What is Multimodal AI?

Definition

Multimodal AI can process and understand multiple types of input — text, images, audio, and video — not just text. Think of it as AI that can see, hear, read, and create.

Traditional AI was text-only. Modern multimodal AI can:

- Analyze photos, documents, and screenshots
- Transcribe and interpret audio recordings
- Understand video content over time
- Generate new images from text descriptions
- Combine information across all of these



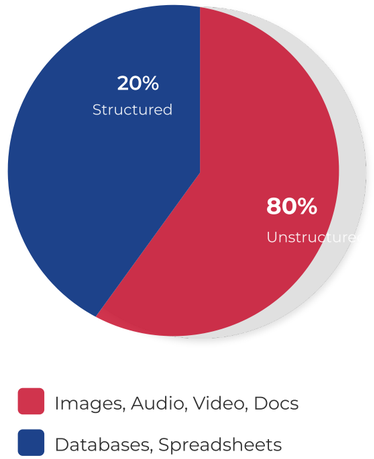
The Business Opportunity

80% of business data is unstructured — and much of it is visual or audio.

Text-only AI leaves enormous value on the table:

- Photos — product images, store displays, receipts, ID documents
- Audio — customer calls, meetings, voicemails, focus groups
- Video — security footage, marketing content, training videos, store walkthroughs

Key Insight: Until recently, this data required expensive specialized software or manual review. Multimodal AI makes it accessible to anyone who can write a prompt.



Real-World Multimodal AI in Action

Warehouse Drones

25x faster counting

Gather AI ($40M Series B, 2026) uses autonomous drones to scan shelves — 99.9% accuracy.

Call Center Analytics

100% calls analyzed

Observe.AI analyzes every call for sentiment, compliance, and coaching opportunities vs. 2-5% manual sampling.

Invoice Processing

68% automation rate

ABBYY and Google Document AI extract structured data from invoices 2-3x faster than manual entry.

Retail Video Analytics

15-30% shrinkage reduction

RetailNext uses overhead cameras to track foot traffic, optimize layouts, and reduce theft.

Manufacturing QC

95-99% defect detection

Landing AI (Andrew Ng) uses cameras to catch defects at rates exceeding manual inspection (industry estimates: 80-90% human accuracy).

Marketing Content

Today's focus

AI analyzes competitor visuals, generates marketing assets, and builds campaign briefs from multimodal data.

Note: Company statistics are illustrative of industry capabilities. Actual results vary by implementation.

Discussion: Multimodal Data in YOUR Industry

Think about your industry or career goal:

- What images does your industry generate or collect?
- What audio data exists (calls, meetings, interviews)?
- What video could AI analyze (walkthroughs, surveillance, demos)?
- How could multimodal AI save time or reveal insights?

Example: Search YouTube for "warehouse drone inventory scanning" to see Gather AI's drones counting inventory 25x faster than humans.

Industries to consider:

- Healthcare: X-rays, pathology slides, patient intake forms
- Real Estate: Property photos, virtual tours, contracts
- Insurance: Damage photos, recorded claims calls
- Education: Lecture recordings, handwritten assignments
- Retail: Store displays, customer interactions, receipts
- Finance: Check images, voice-recorded transactions

Checkpoint: The Multimodal Opportunity

Click an answer to check your understanding.

What percentage of business data is estimated to be unstructured (images, audio, video, documents)?

- About 20%
- About 50%
- About 80%
- About 95%

Section 2

What Can Multimodal AI Actually Do?

Understanding, Generating, and Combining Modalities

Image Understanding: What AI Sees

When you upload an image to Gemini, AI can identify:

- Objects & scenes: "This is a retail store display with seasonal products"
- Text (OCR): Read printed and handwritten text in images (Optical Character Recognition)
- Layout & composition: "The logo is top-center, products are arranged in a grid"
- Context & meaning: "This appears to be a spring promotional display"
- Quality issues: Blurriness, poor lighting, obstructions

Business value: Turn any photo into structured data. A store display photo becomes a competitive analysis. A whiteboard photo becomes meeting notes.

AI Analysis Output:

Products: Spring clothing, accessories
Colors: Pastels, coral, navy
Messaging: "New Arrivals"
Target: Young professionals
Layout: Eye-level focal point

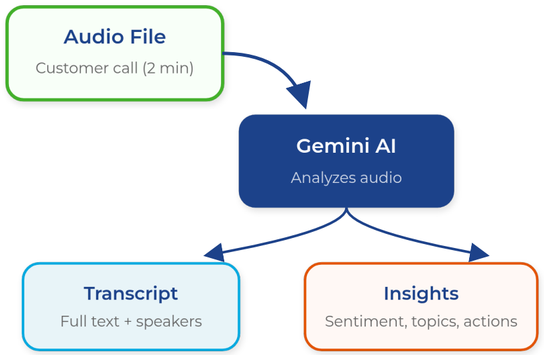
Confidence: High (clear image)


Retail store display for analysis

Audio Understanding: What AI Hears

Upload an audio recording and AI can extract:

- Transcription: Full text of what was said
- Speaker identification: Who said what (by voice pattern)
- Sentiment & tone: Frustrated, satisfied, confused, excited
- Key topics: Product mentions, complaints, compliments
- Action items: What needs follow-up



Unstructured audio → structured business intelligence

Gemini Free Tier

Google Gemini supports audio file uploads on the free tier. Upload short clips (1-3 minutes) for best results.

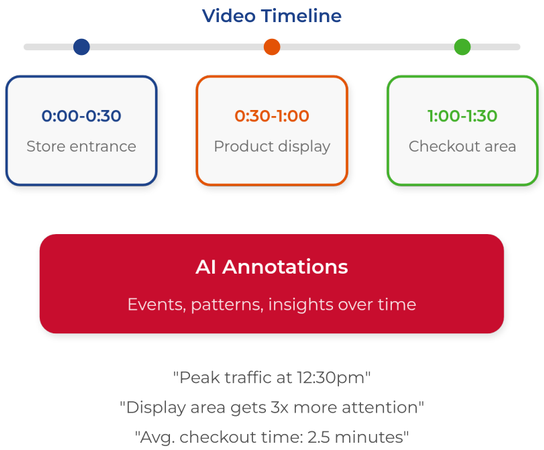
Video Understanding: AI Over Time

Video adds a critical dimension: time. AI can analyze:

- Actions & events: "A customer picks up a product, examines it, puts it back"
- Scene changes: "The video transitions from the entrance to the checkout area"
- Temporal patterns: "Peak traffic occurs at 12:30pm and 5:15pm"
- Combined audio + visual: "The speaker is pointing at a chart while explaining Q3 results"

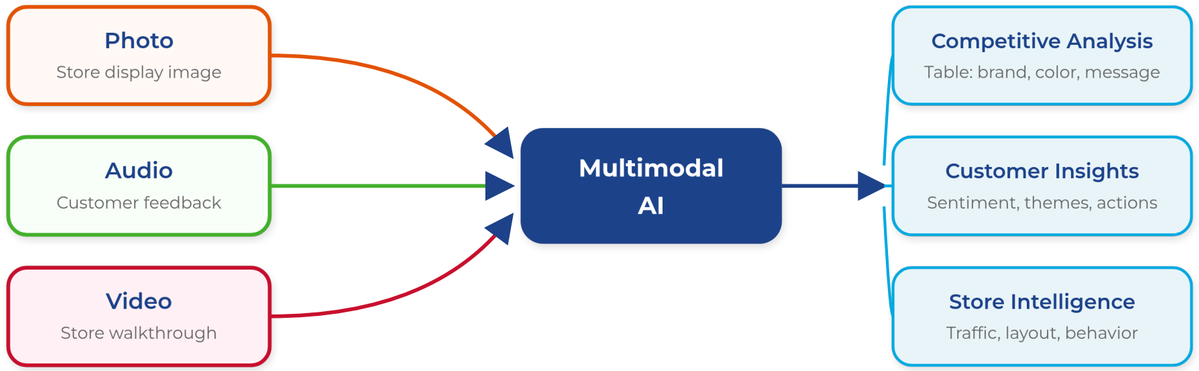
Gemini + YouTube

Gemini can analyze YouTube videos directly — just paste the URL. Free tier supports videos up to 5 minutes; longer videos require AI Pro.



From Unstructured to Structured

The core value of multimodal AI: turning messy real-world data into clean, actionable information.



Day 3 Connection: This is context engineering applied to multimodal data. The same principles — structure your input, specify your output format, iterate — work across all modalities.

Image Generation: AI as Creator

Multimodal AI doesn't just analyze — it can create images from text descriptions.

Business applications for AI image generation:

- Marketing visuals: Social media posts, email headers, ad concepts
- Product mockups: Visualize ideas before production
- Presentation graphics: Custom illustrations for reports
- Concept art: Explore visual directions quickly

Today's Tool

Google Gemini includes image generation — powered by a model called Nano Banana. You can generate images directly in your Gemini chat by describing what you want.

 Text prompt to image generation concept

The promise:

- Speed: Minutes instead of days
- Cost: Free/low-cost vs. designer fees
- Iteration: Unlimited variations
- Accessibility: No design skills needed

The Prompt Matters (Even More)

Remember Day 2? The quality of your prompt determines the quality of the output. For images, this is even more critical.

Vague Prompt

```
"Make a nice ad for a store"
```

Result: Generic, bland image with no brand identity, wrong colors, unclear message.

Detailed Prompt

```
"Professional spring marketing banner for Beacon Retail:
bright pastel colors, clean modern layout, seasonal
products on white shelving, warm natural lighting,
aspirational lifestyle feel, 'Spring Collection' text"
```

Result: On-brand, polished, ready-to-use marketing asset.


Before and after prompt comparison

Day 2 Callback: The RCTFC framework works for images too. Role (professional photographer), Context (spring marketing campaign), Task (create a banner), Format (16:9, bright, clean), Constraints (brand colors, no text errors).

Current Limitations: An Honest Assessment

AI CAN Reliably:

- Describe what's in an image with high accuracy
- Transcribe clear audio recordings
- Summarize video content and key moments
- Generate creative marketing visuals
- Extract structured data from photos
- Identify objects, scenes, and text

AI CANNOT Reliably:

- Identify specific people in photos (a deliberate safety restriction)
- Generate perfectly readable text in images
- Guarantee factual accuracy in generated images
- Understand deeply ambiguous or ironic content
- Create pixel-perfect brand-compliant designs
- Replace professional designers for final production

Trust But Verify

AI-generated images may contain hallucinated details (wrong number of fingers, misspelled text, impossible physics). Always review AI output before using it professionally. The iteration skill — prompt, review, refine — is what separates good results from great ones.

Checkpoint: Capabilities & Limitations

Click an answer to check.

Beacon wants to generate marketing images with AI. Which of these is a known limitation they should plan for?

- AI cannot generate images of products
- AI may misspell text embedded in generated images
- AI-generated images are always low resolution
- AI can only generate black-and-white images

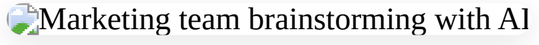
Section 3

Building Beacon's Spring Campaign

Applying Multimodal AI to a Real Marketing Workflow

The Scenario: Beacon's Spring Campaign

Mission: Beacon Retail Group is launching a spring marketing campaign. Your team will use multimodal AI to research, create, and plan — all in one session.



Your 4-step workflow:

1. Analyze competitor marketing images
2. Listen to customer feedback audio
3. Generate marketing visuals with AI
4. Build a structured campaign brief

This is the same workflow real marketing teams use — but with AI accelerating every step.

Step 1: Competitive Visual Analysis

Upload competitor marketing images to Gemini and extract structured competitive intelligence.

```
Prompt:
"Analyze this marketing image. Identify: target audience, color palette, key messaging, product positioning, and emotional appeal. Present as a structured table."
```

From a single image, AI can identify:

- Color palettes and brand positioning
- Target audience signals
- Messaging themes and emotional hooks
- Layout and design patterns

Structured Output:

Target	Young professionals, 25-35
Colors	Coral, white, gold
Message	"Fresh starts, new style"
Position	Aspirational lifestyle
Appeal	Optimism, renewal

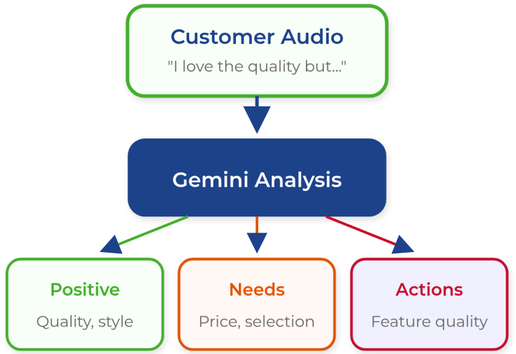
Lab preview: You'll do this with 2-3 real competitor images.

Step 2: Customer Voice Analysis

Upload customer feedback audio to Gemini. AI extracts marketing-relevant insights from the recording.

- What customers love about Beacon products
- What frustrates them (unmet needs = opportunity)
- Language they use (messaging inspiration)
- Specific product mentions (what to feature)

Day 3 Connection: Use structured formats from Day 3 to organize the audio insights into a clean table: Theme | Sentiment | Quote | Marketing Action.



Audio → Structured marketing insights

Step 3: Generate Marketing Visuals

Now the creative part: use AI to generate marketing images for Beacon's spring campaign.

```
Example Prompt:

"Professional spring marketing image for Beacon Retail: bright,
airy store interior with seasonal products on modern white
shelving, warm natural lighting, pastel accent colors,
aspirational lifestyle feel, clean modern composition"
```

The iteration process:

1. Write your first prompt (start broad)
2. Review the result — what's good? What's off?
3. Refine the prompt with specific adjustments
4. Repeat until you're satisfied (min 3 iterations)

 Generated Beacon spring campaign visual

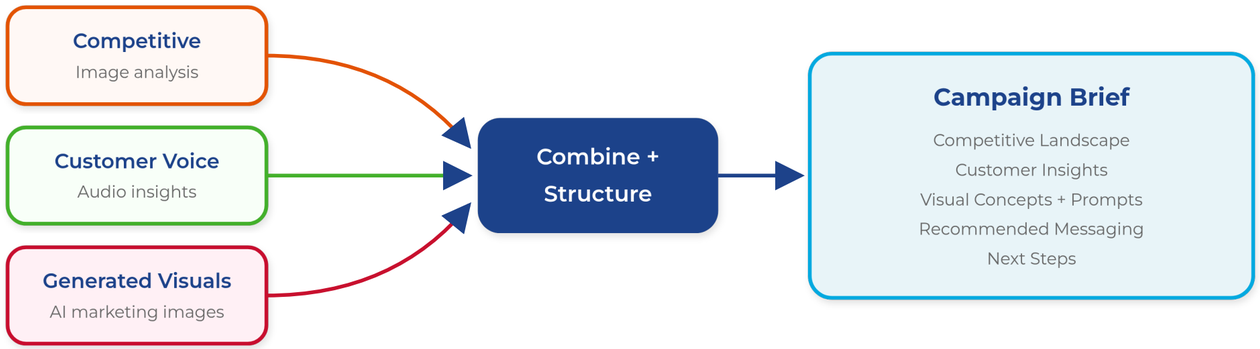
Campaign Suite (Lab):

- Email newsletter hero image
- Social media post graphic
- In-store display concept

Each requires a different prompt tailored to the channel.

Step 4: Structured Campaign Brief

Combine all your multimodal insights into a single structured deliverable.



This is multimodal context engineering. You've taken images, audio, and AI-generated visuals — combined them with structured formats from Day 3 — and produced a professional marketing deliverable. That's the workflow.

Human-in-the-Loop: Why AI Needs You

AI generates drafts. Humans curate and refine. Here's why marketing professionals still matter:

- Brand voice: AI doesn't know your brand's personality the way you do
- Cultural sensitivity: AI can miss context that matters to your audience
- Strategic alignment: Does this support the campaign's actual goals?
- Legal & compliance: Copyright, trademark, claims review
- Quality bar: Is this good enough for your customers to see?

AI accelerates; humans ensure quality

Ethical Considerations for Multimodal AI

With the power to analyze and create visual content comes responsibility. Here's what business leaders must consider:

⚖️ Copyright & Ownership

AI-generated images may inadvertently resemble copyrighted work. Always verify originality before publishing. Copyright law for AI-generated content is still evolving.

📢 Disclosure & Transparency

When should you tell customers that content was AI-generated? Best practice: always disclose, especially in marketing, journalism, and customer communications.

🛡️ Deepfakes & Misuse

AI that creates realistic images can also create misleading ones. This is why models restrict face generation and include safety filters — it's a feature, not a bug.

🧬 Bias in Visual Content

AI-generated images can reflect biases in training data — certain demographics, body types, or cultural norms may be over- or under-represented. Review with diverse perspectives.

Beacon's policy: All AI-generated marketing content must be reviewed by a human for brand alignment, cultural sensitivity, and accuracy before publication. Include "AI-assisted" attribution on generated visuals.

ROI: The Business Case for AI-Powered Marketing

10x
Faster concept art
Minutes vs. days

70%
Cost reduction
AI mockups + designer polish

50+
Variations per hour
vs. 2-3 manually

Beacon's math:

- Before: $5,000/campaign for a designer, 10 images, 2 weeks
- After: AI generates 50 concepts in 2 hours, designer polishes best 10 for $1,500
- Savings: $3,500/campaign + 12 days faster to market

Vendor ROI Claims

Be cautious with vendor-published ROI figures. They often reflect ideal conditions. Your actual savings depend on campaign complexity, brand requirements, and how much human review is needed. Calculate YOUR numbers.

Section 4

Putting It All Together

Beyond Marketing: Where Else?

Marketing was our deep dive — but the same multimodal principles work everywhere.

Operations

Drone inventory → stock analysis. Visual QC → defect reports.

Customer Service

Call recordings → sentiment trends. Chat logs + screenshots → issue resolution.

Finance

Invoice photos → structured data. Receipt images → expense reports.

HR & Training

Interview recordings → structured notes. Training videos → knowledge checks.

Facilities

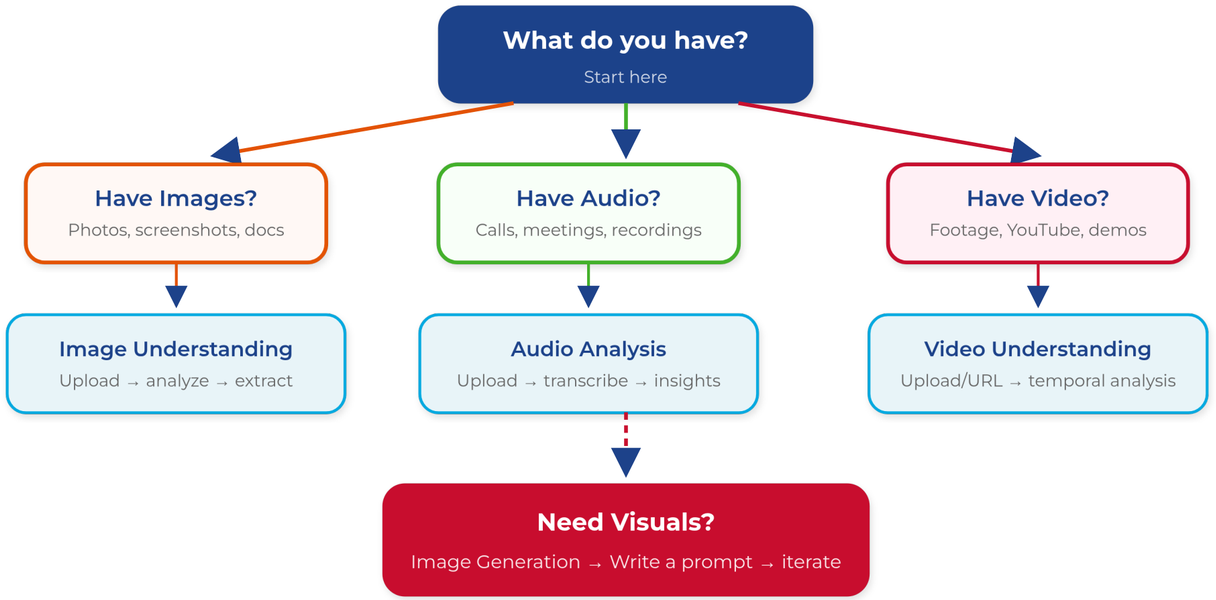
Store walkthrough videos → compliance audits. Photo inspections → maintenance tickets.

Product Development

Customer review audio → feature priorities. Competitor product photos → design analysis.

Pattern: Multimodal input (image, audio, video) + AI + structured output format = actionable business intelligence. The modality changes; the workflow stays the same.

The Multimodal Toolkit: When to Use What



Key Takeaways

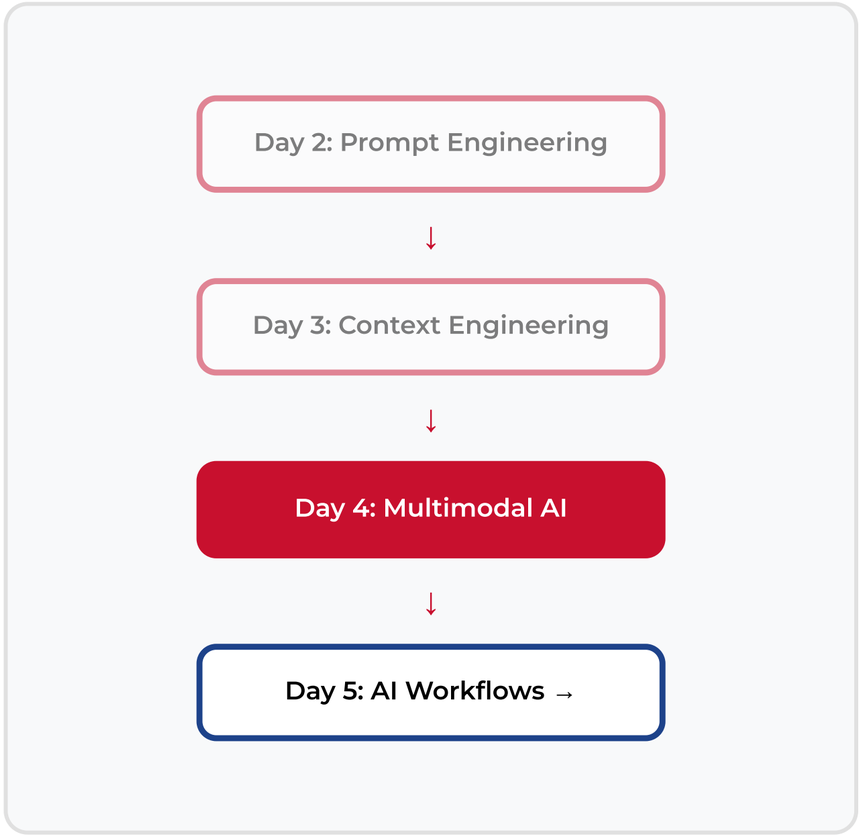
1. Multimodal AI processes images, audio, and video — not just text. It can see, hear, and create.
2. The core value is turning unstructured data into structured insights. Photo → analysis table. Audio → sentiment report. Video → timeline.
3. AI can generate images, not just analyze them. Marketing visuals, mockups, and concepts — all from text prompts. The prompt quality determines the result quality.
4. Human-in-the-loop is essential. AI generates drafts; humans ensure brand consistency, cultural sensitivity, and strategic alignment.
5. The multimodal workflow extends Day 3's context engineering. Same principles (structure, iterate, verify) — now applied to images, audio, and video.

What's Next

Day 5 Preview

Building on everything so far: prompts, context engineering, and multimodal AI — combined into real business workflows.

- Putting it all together with AI Studio
- Building end-to-end business workflows
- From individual skills to integrated solutions



Questions?

Let's discuss before moving to the lab.

Up next: Beacon's Spring Marketing Campaign Lab

You'll analyze competitors, listen to customers, and generate marketing visuals — all with AI.