SAMIR GENAI ACADEMY

PROFESSIONAL STUDY GUIDE SERIES

Phase 01 · GenAI Foundations

04

Multimodal AI Systems

Vision, audio, video & code — unified model architectures

Module 04 of 25 · **1 hour**

Instructor: **Samir Kumar** · Edition 2026

STUDY GUIDE

01 Overview

> Vision, audio, video & code — unified model architectures. This guide gives you the core theory, the tools that matter, a hands-on build, working code, and the pitfalls to avoid — everything you need to learn the topic to a professional standard and apply it in real projects.

1 hour	4 + 4	4	04/25
LESSON LENGTH	CONCEPTS + BUILD STEPS	FRAMEWORKS & TOOLS	PHASE 01

02 Learning Objectives

- Understand how models fuse text, image, audio, and video into one representation.
- Build image-analysis and OCR workflows.
- Add speech-to-text and text-to-speech to applications.
- Compose multi-step multimodal pipelines.

03 Core Concepts

Unified representations

Multimodal models encode each modality (pixels, audio frames, text) into a shared embedding space so the model can reason across them — describe an image, answer questions about a chart, or transcribe and summarise audio in one pass.

Vision & OCR

Vision-language models read documents, diagrams, screenshots and photos. They handle layout-aware OCR far better than classic engines for messy real-world images, and can answer questions grounded in the visual content.

Speech pipelines

Whisper-class models transcribe audio robustly across accents and noise; TTS models synthesise natural speech. Chaining ASR -> LLM -> TTS yields voice assistants and meeting summarisers.

Composing workflows

Real value comes from pipelines: screenshot -> extract -> reason -> act; or audio -> transcript -> summary -> email. Treat each modality step as a tool the orchestration layer calls.

04 Visual Diagram

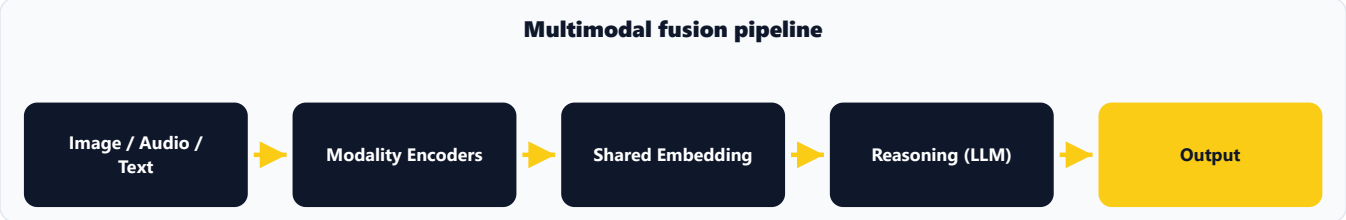

Figure 4 — Each modality is encoded into one shared space so the model can reason across image, audio, and text together.

05 Key Frameworks & Tools

Whisper | Vision-capable Claude/GPT/Gemini | Pillow / OpenCV | ffmpeg

06 Hands-On Build

🛠 Image Analysis OCR + Voice Synthesis with Whisper + Multimodal Workflow

1. Send an image to a vision model and extract structured fields (invoice/receipt).
2. Transcribe an audio clip with Whisper and summarise the transcript.
3. Generate spoken audio from the summary with a TTS model.
4. Chain image -> extraction -> reasoning -> spoken result into one script.

07 Code Walkthrough

Image analysis with a vision model — PYTHON

```
import base64, os
from anthropic import Anthropic

client = Anthropic(api_key=os.environ["ANTHROPIC_API_KEY"])
img = base64.standard_b64encode(open("invoice.png", "rb").read()).decode()

resp = client.messages.create(
    model="claude-opus-4-8", max_tokens=512,
    messages=[{"role": "user", "content": [
        {"type": "image", "source": {"type": "base64",
         "media_type": "image/png", "data": img}},
        {"type": "text", "text": "Extract invoice number, date and total as JSON."}]}],
)
print(resp.content[0].text)
```

08 🧪 Lab Assistant

Lab 4 · Invoice OCR with a vision model

Prerequisites:A vision-capable API key · A sample invoice.png

1. **Base64-encode the image and send with an extraction instruction**
   > Expected: JSON with invoice no., date, total
2. **Resize the image to ~1024px wide and re-run**
   > Expected: Similar accuracy, lower cost/latency
3. **Add a validation check on the total field**
   > Expected: Flags malformed values before downstream use

Troubleshooting

- Huge/slow request → downscale the image first.
- Wrong fields → make the schema explicit in the prompt.

09 Pitfalls & Key Takeaways

⚠ COMMON PITFALLS

- Sending huge images uncompressed — resize first to cut cost and latency.
- Trusting OCR output without validation for critical fields.
- Forgetting audio format/sample-rate requirements for ASR models.

✓ KEY TAKEAWAYS

- Multimodal models share one embedding space across modalities.
- Vision models beat classic OCR on messy real-world documents.
- Build value by chaining modality steps into pipelines.

📺 **Companion video:** Watch the module 04 lesson in your Student Dashboard, then use this guide to revise, copy the code, and complete the build. Download both for offline study.

10 Further Resources

OpenAI Whisper repo • Vision API docs (Claude/GPT/Gemini) • ffmpeg documentation