GENERATIVE AI · TAUGHT WITH GOOGLE GEMINI

Multimodal AI

Models that read documents, charts, screens and video. You build a real extraction pipeline, then learn the specific ways vision models invent table cells and miss small print.

Advanced | 5 phases | 26 sessions | 39 contact hours

Syllabus

5 phases, 26 sessions. Each phase ends in something you have built.

1 Pixels Into Tokens

Know what an image actually costs and what the model sees after your file is resized.

5 sessions
$87 alone

1.1 What The Model Sees

Vision encoder, MLP merger, and why a 4000px screenshot is downsampled before anything reads it.

FREE PREVIEW

1.2 Counting Visual Tokens

Gemini's 258 tokens per PDF page against Claude's 28x28 patch formula, priced over a thousand pages.

1.3 Native Resolution & M-RoPE

Dynamic resolution, NaFlex, and the 2x2 patch merge that turns a 224x224 image into 66 tokens.

1.4 Current, Legacy, Retired

Pixtral, Llama 3.2 Vision and Qwen2.5-VL are legacy, and video support rarely means a video file.

1.5 Benchmarks After Saturation

MMMU published its test answers in February 2026, DocVQA sits at ceiling, and what to read instead.

By the end: **A token and cost model for your own images, PDFs and video, checked against three provider APIs.**

2 Documents, Honestly

Pull structured data out of real PDFs and know exactly which fields are wrong.

6 sessions
$87 alone

2.1	**The Eval Set First**	Fifty pages from your own corpus, labeled by hand, before you choose a single model.
2.2	**OCR Or Native VLM**	Chandra, olmOCR 2 and a 0.9B PaddleOCR-VL against Gemini 3.6 Flash on the same pages.
2.3	**Tables That Lie**	Merged cells, spanning headers, and the row a model invents because the pattern wanted one.
2.4	**Charts Without Numbers**	Reading values off an unlabeled axis, and the chart questions frontier models still get wrong.
2.5	**Schema-Constrained Extraction**	JSON schemas, per-field nulls, and making the model decline a field that is not on the page.
2.6	**Reading Order & Repetition**	Multi-column layout, footnotes, and the decode loop that repeats one paragraph until you stop it.

By the end: **An extraction pipeline over 200 real pages with a per-field accuracy table and a cost per page.**

3 Grounding & Retrieval

Make the model point at what it read, then find the right page out of ten thousand.

5 sessions
$87 alone

3.1	**Boxes, Points, Masks**	Gemini returns ymin first, Qwen3-VL returns xmin first, and that swap ruins every overlay you draw.
3.2	**Cite The Pixels**	Every extracted field carries the box it came from, so a human can verify a page in seconds.
3.3	**Multimodal Embeddings**	SigLIP 2, Gemini Embedding 2 and voyage-multimodal-3.5, scored on your pages rather than on ViDoRe.
3.4	**Late Interaction, Costed**	ColPali-style retrieval wins recall and costs 100x to 1000x more vectors per page to store.
3.5	**Cross-Modal Search**	One index over pages, frames and transcripts, and the queries where plain keyword search still wins.

By the end: **A visual search service returning the page, the box on it, and a measured recall@5.**

4 Video That Fits

Answer questions about two hours of video without paying for two hours of frames.

5 sessions
$87 alone

4.1 Frames, FPS, Budget
Gemini samples at 1 FPS for about 300 tokens a second, so a 1M context buys roughly one hour.

4.2 Sampling That Keeps Evidence
Uniform sampling against keyframe and query-guided selection, scored on the same question set.

4.3 Audio Is Half The Answer
Transcript alongside frames, and the lecture questions no amount of frame sampling will answer.

4.4 Long Video Needs Retrieval
Chunk into clips, retrieve, then answer, because LVBench tops out near 64 against 94 for humans.

4.5 Temporal Failure Modes
Event ordering, counting repeats, and needle-in-a-haystack search across two hours of footage.

By the end: **A video QA tool over a two-hour recording that cites a timestamp for every answer.**

5 Screens & Rails

Drive a real interface from screenshots, and stop a screenshot from driving you.

5 sessions
$87 alone

5.1 Screenshots As Input
High-resolution displays, tiny targets, and why ScreenSpot-Pro still separates models that tie.

5.2 Acting On A Screen
Gemini 3.6 Flash built-in computer use against Anthropic's computer-use tool on ten identical tasks.

5.3 OSWorld Is Not A Number
85 percent, 83 percent, a 72.36 human baseline, and three harnesses that do not compare.

5.4 Injection Through Pixels
Typographic and steganographic instructions hidden in a screenshot, and defenses that partly hold.

5.5 Approve Or Decline
Human sign-off on irreversible actions, least privilege, and the automation you refuse to build.

By the end: **A computer-use agent scored on a task set you wrote, plus an image-injection red-team report.**

Tools you will use

Google Gemini 3.6 Flash | Claude Sonnet 5 | Qwen3-VL 235B-A22B | olmOCR 2 | PaddleOCR-VL | ColQwen2.5 | Qdrant | SGLang | torchcodec | Holo3 35B-A3B

What you will build

1. **The Page Ledger**
   Two hundred real pages, per-field accuracy, cost per page
2. **Find The Frame**
   Ten thousand pages and two hours of video in one index
3. **The Screen Agent**
   Your own task set, a real score, an injection red team

Registration

Phase 1: Pixels Into Tokens	$87
Phase 2: Documents, Honestly	$87
Phase 3: Grounding & Retrieval	$87
Phase 4: Video That Fits	$87
Phase 5: Screens & Rails	$87
Whole track, all 5 phases	**$360**

Buying every phase separately costs $435, so the whole track saves $75.

No payment is taken online. Registering creates an order and payment instructions are sent with it. Access opens once payment is confirmed.