M-Lab CARTE AI Workshop 2025

Beyond Text – Multimodal AI

What is multimodal AI?

- AI that works across different types of data
- Today we will cover image, video and audio
- All of these depend on LLMs!
  - This was not always the case
- LLMs help us to turn a description of what we want into meaningful information
- But the other parts of the model are also transformer-based

Why multimodal matters now

- All of these applications have existed for some time
  - In particular, audio – text-to-speech has been around for decades
- Transformer-based models have rapidly advanced capabilities
- As LLMs slow in performance improvement, multimodal remains a key frontier for development

Image Generation

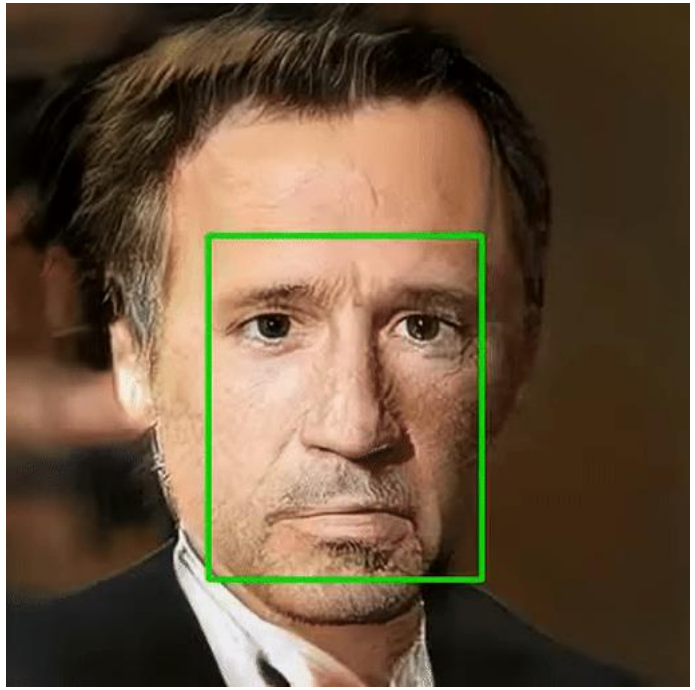
How image generation works

- Vision *understanding* was key technology of the 2010s
- Challenge: generating new imagery
- First breakthrough: Nvidia "GAN" models
  - Generative Adversarial Network
- Adversarial: can another network tell the difference between real & fake?

2014 2018

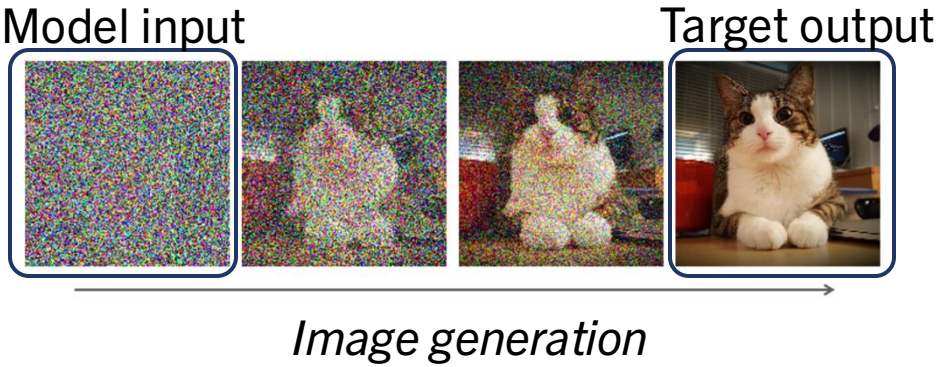
Diffusion models

- GAN models create convincing imagery, but output cannot be directly controlled
  - *Language* understanding is missing!



Diffusion models

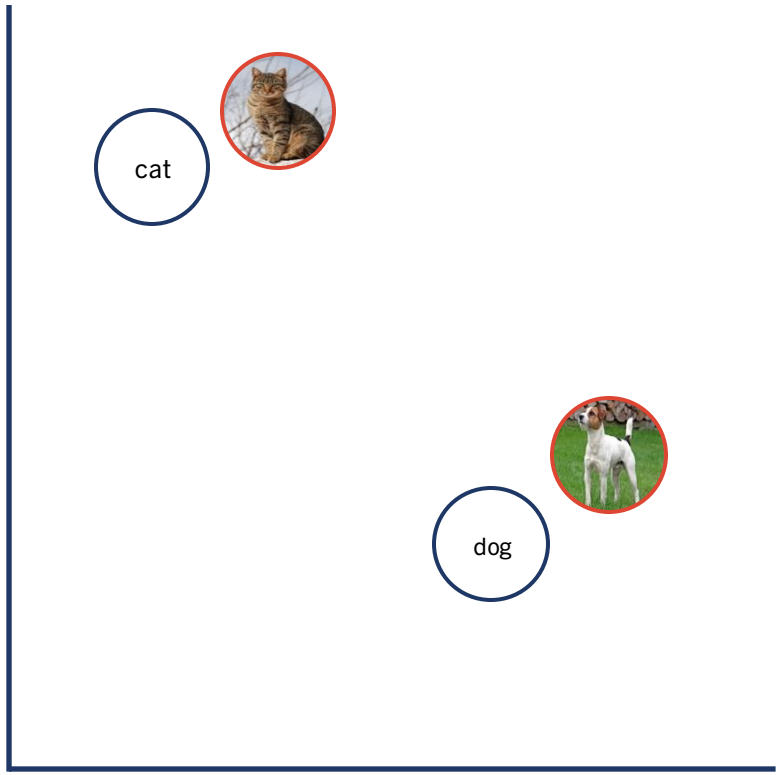
- Image generation models started out as image *denoising* models
- Training process: take a regular image, add noise, then train by comparing to the original
- Insight: what if we make the input *all noise,* but teach it to predict real content?



https://poloclub.github.io/diffusion-explainer/

Text-to-image

- With the advent of large language models, it soon became possible to direct the generation using text description
- This is achieved by representing text and imagery in the same *embedding space* as one another
- Think back to the “Marilyn Monroe neuron” – text and imagery representing the same concept

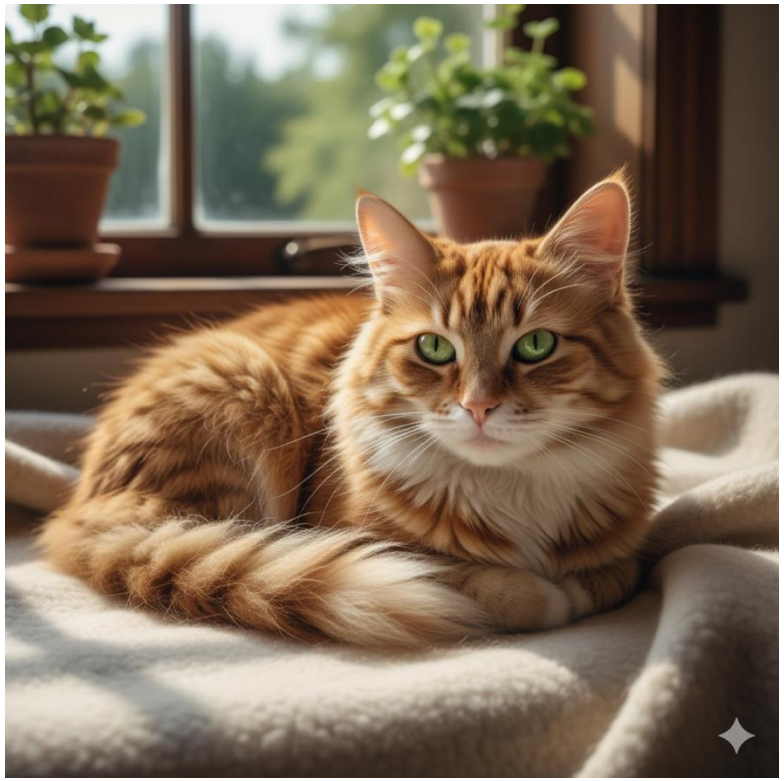


Prompting for images

- When requesting an image, specificity is critical
- Entering “cat” could produce a drawing, photo, cartoon…
- Specificity also allows for *consistency*
- Negative prompting allows you to enter words that the image should *not* look like
- LLMs can help to expand on a prompt based on ideas, example imagery

Prompting for images

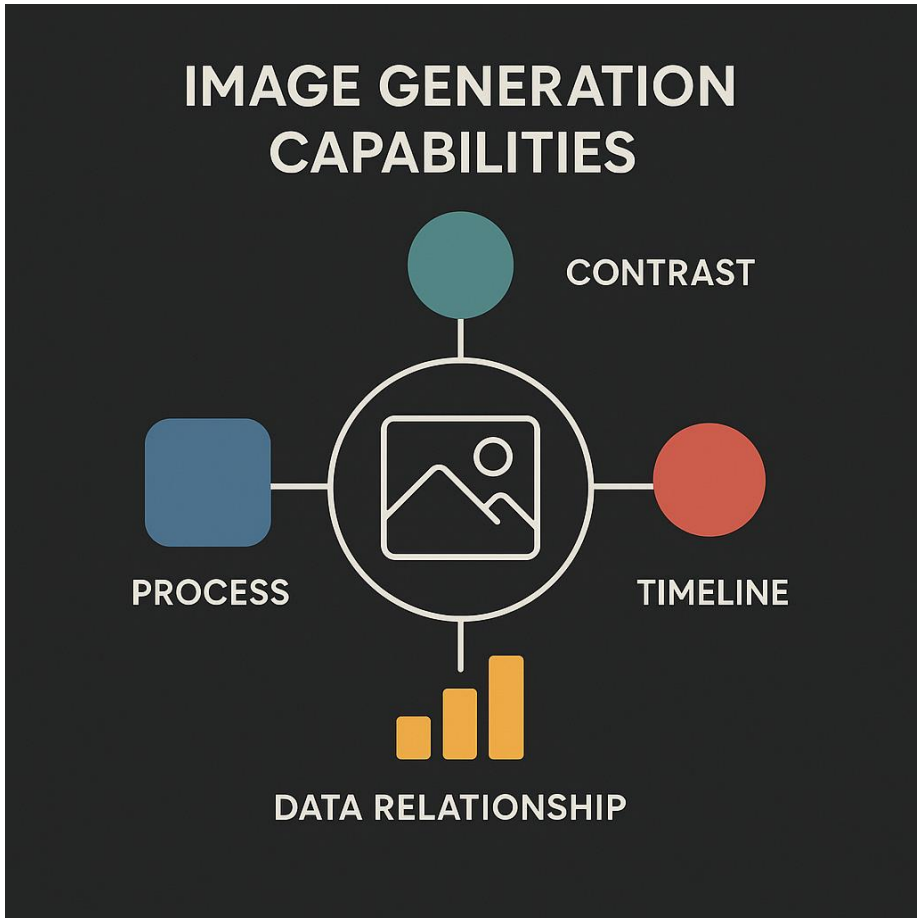
“cat”



A candid outdoor photograph of a cat perched on a weathered wooden fence during light rain. The cat’s fur is slightly damp, its ears alert, and raindrops are visible mid-air against a blurred background of green leaves. Natural overcast lighting gives soft, even tones. Shallow depth of field with realistic focus falloff. Shot from eye level with the cat using a 35 mm lens equivalent, capturing lifelike texture and atmosphere — no stylization or artificial lighting.

Key Players - OpenAI

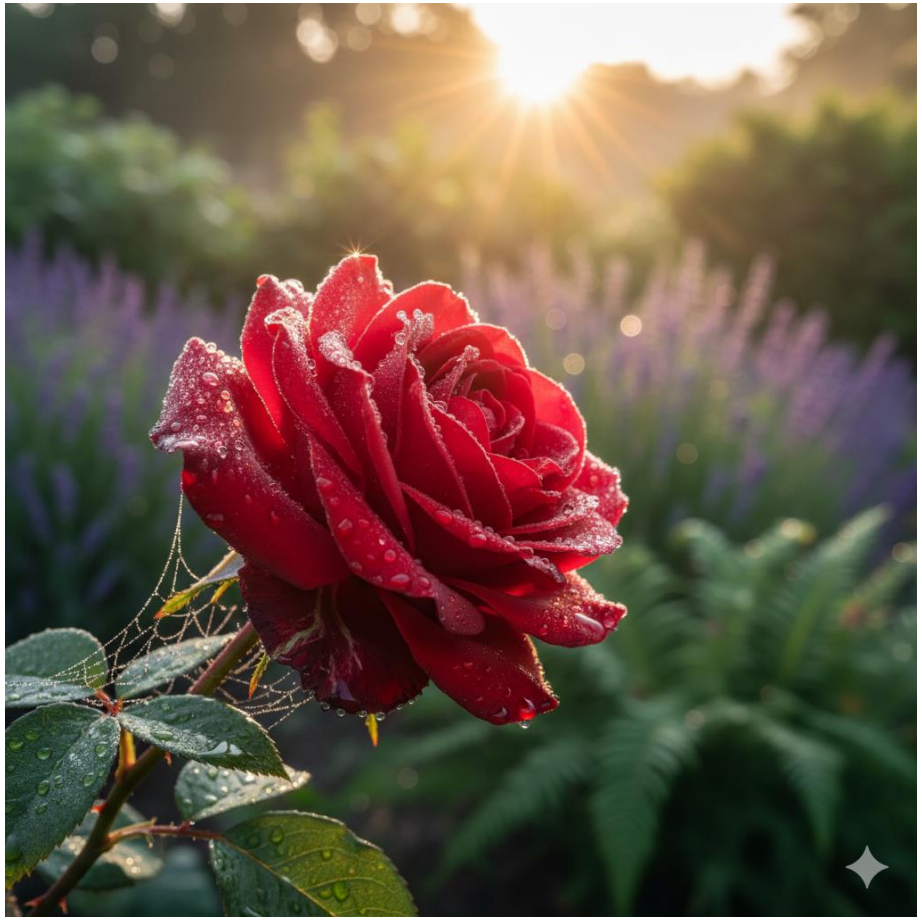
- Text-to-image improved dramatically with diffusion models 2022-2023. DALL-E 3 brought major leap in prompt understanding
- Natural language prompts work well without “prompt engineering” tricks
- Integrated directly into ChatGPT
- Handles text within images better than most competitors
- Commercial rights included automatically



Created by DALL-E 3

Key Players - Google

- Highest photorealism in benchmarks
- Superior prompt understanding
- Exceptional detail and lighting effects (world understanding)
- Wide range of art styles
- Image editing functionality



Created by Gemini 2.5

Other Players

- Midjourney
  - Artistic gold standard, especially for stylized work
  - Discord-based interface (steep learning curve)
  - Strong aesthetic coherence (consistency)
- Stable Diffusion
  - Open source, can be run on-premises
  - High degree of customization
  - Inexpensive API access and free local usage
- Adobe Firefly
  - Commercially safe – only trained on fully licensed content
  - Integrated into Adobe Creative Suite

Use cases

- Still not “production ready” – but getting there quickly
- Useful for developmental stages
  - Prototyping
  - Concept art
  - Product visualization
- Allows non-artists to quickly visualize ideas

Limitations

- Text coherence: rapidly improving but still prone to errors or illegible sections
- Anatomy: humans are highly sensitive to subtle errors (”uncanny valley” effect)
- Consistency: difficult to generate exactly the same detail each time
- Copyright concerns

Audio Generation

Overview

- Speech
  - By far the area of greatest focus
  - Significant strides even in the pre-transformer era
  - Increasingly natural
- Music
  - Complex for models: musical “ideas” must remain consistent through a piece
- Sound effects
  - Perhaps least complex but similar “uncanny valley” effect to faces

Text-to-Speech

- Text-to-Speech (TTS) has existed for decades
- Typical:
  - Break text into standard word sounds (phonemes)
  - Combine pre-created (or even pre-recorded) phonemes
  - Result: speech
  - Only the first step requires prediction – phonemes vary depending on word order, context, etc. (e.g. “I will read” vs “I have read”)

Microsoft SAM (1982)
rule-based signal generation

DECtalk (1990s)
phoneme rules + prosody control

Google WaveNet (2016)
deep-learning based

Modern TTS

- State-of-the-art text to speech works similarly to image generation
- Text is *co-embedded* with *audio tokens*
- A denoising model then converts audio noise into coherent speech using the *audio tokens* as input
- Tone, emotion and pacing are encoded implicitly in the process
  - One phoneme may have many corresponding audio tokens depending on these additional features
- In this way, the model predicts more than a flat computer voice

Google NotebookLM (2025)

Voice cloning

- Voice cloning adds a *speaker identity* dimension to text-to-speech
- During training, the model learns to extract vocal characteristics: pitch range, timbre, accent, speech patterns
- The model then generates speech that matches both the *text content* and the *voice identity*
- Modern systems can clone a voice from as little as 3-10 seconds of audio
- Quality improves with longer reference samples (1-2 minutes gives near-perfect cloning)

Music generation

- Text-to-music models generate audio tokens sequentially, like speech models
- But music requires coordinating multiple elements: melody, harmony, rhythm, structure
- Models are trained on millions of songs to learn patterns like chord progressions, verse-chorus structure, and genre conventions
- The model generates audio tokens that satisfy both the musical structure (coherent song) and your constraints (style, mood, tempo)

Video Generation

Video challenges

- At their most basic, video generation can be achieved by training image generators to slightly vary content
  - Therefore producing frames
- However, this approach lacks a “plan” of what should happen in the video
- Even worse, there is no understanding of what has already occurred

StableDiffusion (2023)

How video generation works

1. Text description is converted into embeddings
2. *Spatio-temporal chunks* (e.g. four seconds of top-right corner) initialized
3. Content of each chunk is predicted based on input *(across time)*
4. Model evaluates “plan” all together — checks for coherence
5. Frames are generated from plan using diffusion

https://generative-animation-explainer-871047044699.us-west1.run.app/

Key Players - OpenAI

- Sora 2 released September 30
- Emphasis on *world understanding* – particularly physics
- Sora 2 phone app – TikTok for AI generated videos
- Raises significant questions about corporate responsibility

Sora 2 (OpenAI)

Key Players - Google

- Veo 3 launched mid-2025
- Emphasis on *integrated audio generation*
- Enables the generation of synchronized sound effects, dialogue and lip sync
- Strong physics understanding and cinematic control
- Integrated into Google's cloud platform

Veo 3 (Google)

Other players

- Runway Gen-4
  - Professional creative suite
  - Text-to-video, image-to-video, video editing tools
  - Higher learning curve and cost
  - Partnered with IMAX, Lionsgate
- Pika Labs
  - Fast generation, user-friendly interface
  - Quick iterations for social media and prototyping

Limitations

- Short duration
- Physics errors
- Consistency
- Complex motion
- Editing control
- Text and fine details

Putting it Together

Cross-Modal Applications

What is Cross-Modal?

- Models that understand and generate across multiple modalities
- One-stop-shop to produce a variety of content
- Aim is to eventually produce content automatically end-to-end

Vision + Language

- Vision-Language Models (VLMs) offer capability beyond generation
- Image captioning
- Visual Question Answering (VQA)
- Visual Search

GPT-5

Gemini

Multimodal Content Pipelines

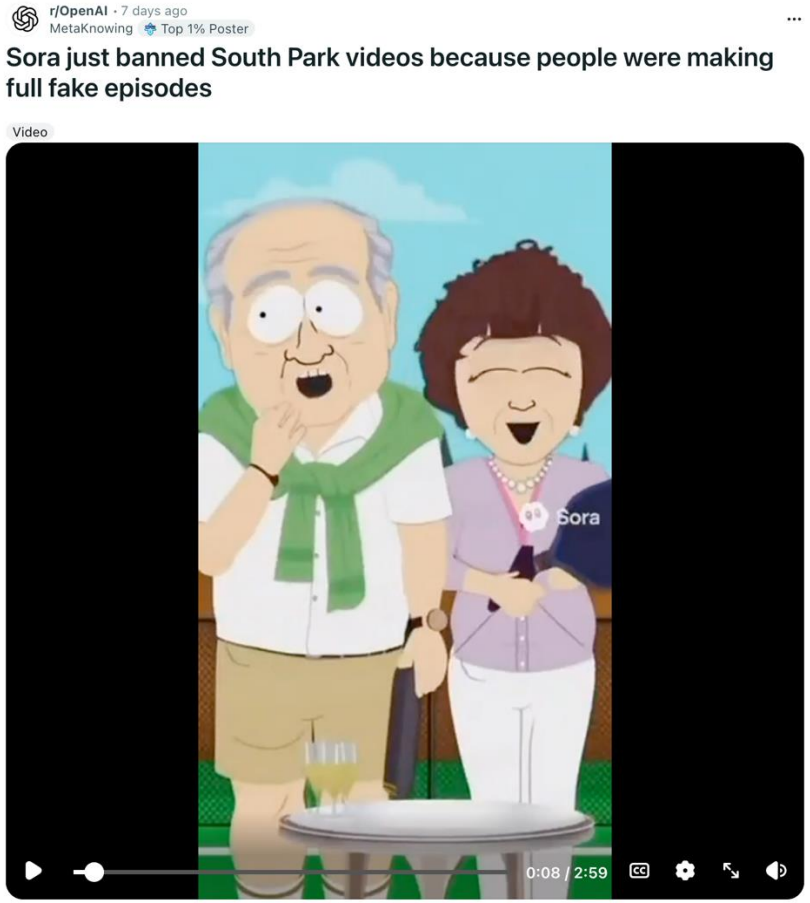
- Aim is to assist (or handle entirely) every step of the process
- Script
- Concept art
- Video
- Voiceover
- Technically feasible now, but creativity is a factor

Quality and Consistency Challenges

- Models struggle to produce content that retains visual consistency
- When samples are short, this challenge is made harder
- Improvements have been made, but humans are good at noticing subtle variation
- Full-length films often promised but yet to meaningfully appear

Practical considerations

- Models are slow to run and expensive
- Sora 2: $0.50/second
  - Prohibitive for quick iteration and prototyping
- However – a modern Hollywood film could cost $9,000/second
  - This is the rival companies like OpenAI and Google are aiming for
- Rights management is a concern
- Disclosure over use of AI?



https://www.reddit.com/r/OpenAI/comments/1nyl15y/sora_just_banned_south_park_videos_because_people/

What’s next

- Longer videos
- Real-time generation
- Consistency
- Agentic orchestration…?

Questions?