Where is this all going?

D.A. Forsyth, University of Illinois at Urbana Champaign

SOME APPLICATION SLIDES MISSING

Budding topics

- Generating images, video, etc
- The merging of topics in high-compute regimes
- Divergence between low-compute and high-compute methods

Main points

- You can generate images, video, etc. from random variables
- You can train generators without explicit labelling
- Generators are intriguing
  - They seem to know stuff that isn't in the training data
  - They seem not to know stuff that is in the training data
- The technology is very poorly understood
  - and is fantastically hard to evaluate

Images from random numbers

- Astonishing, recent fact:
  - We can generate images from random numbers
    - and conditioning information
- Two major procedures:
  - Generative adversarial network
  - Diffusion methods

In Limine: I'm ignoring text/image links, editing, etc.

to talk about scene representation and understanding

Applications: make commercial art cheaper

and political discourse nastier

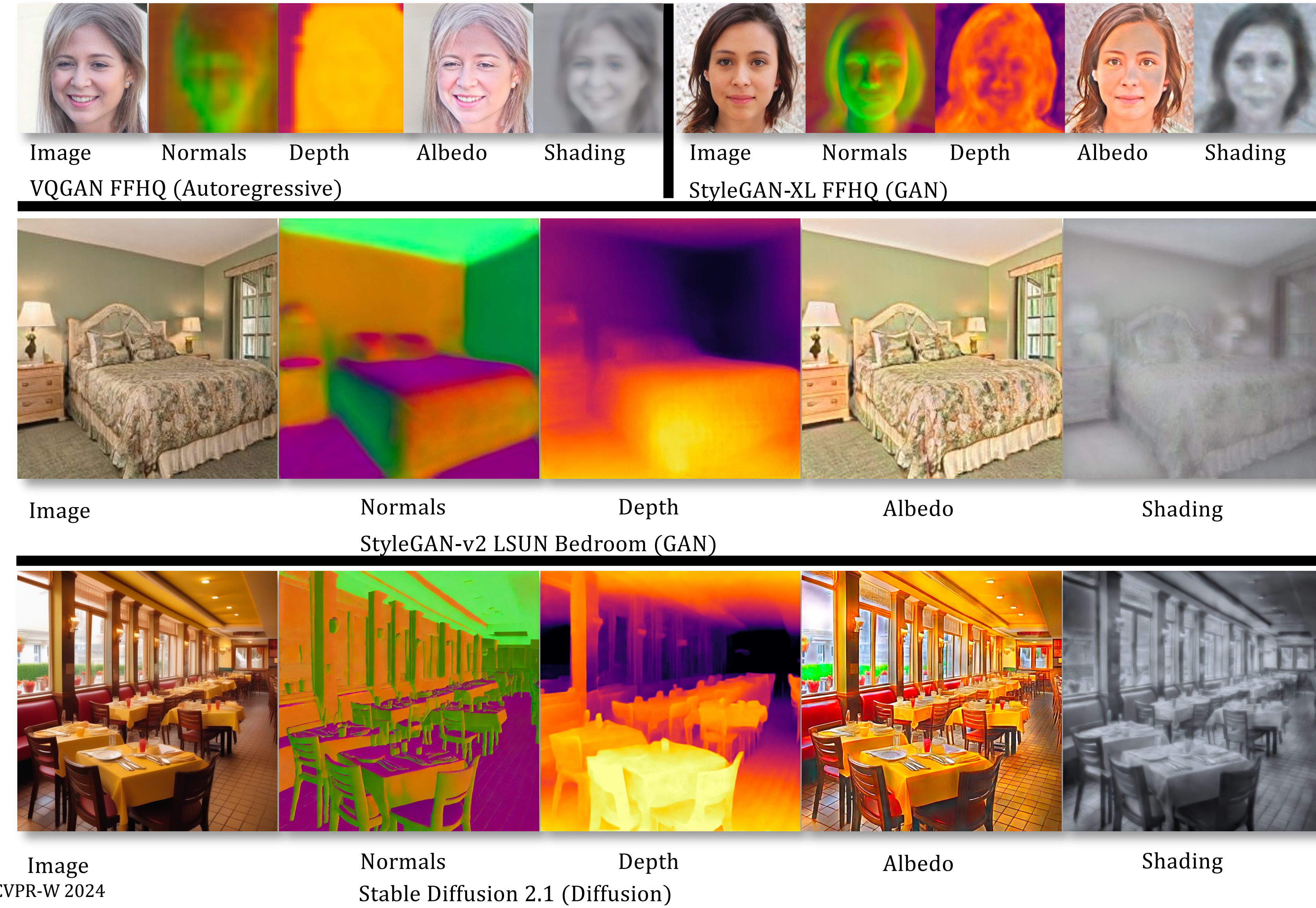
Articles of faith

- There is a probability distribution on all images
- A diffusion model produces a fair sample from that distribution

- Articles of faith because:
  - they’re believed in the face of inconvenient facts
    - obvious falsification below
    - you can’t do two-sample testing in high dimensions and get the answer right
    - metrics:
      - absurd (FID);
      - ineffectual (ask people);
      - inaccurate (ask VLMs)
  - bad things might happen to unbelievers

So why bother with image generators?

- Silly arguments
  - You can get rich by taking money from commercial artists
    - (who haven’t got any)
  - Prurience
    - The NSFW world just *loves* image generators
      - fine-tuned specialized generators are traded on a large scale
- Generators know a lot about pictures, and data is easy
  - Surprise: They know a lot about relighting, intrinsics, etc.
      - without being trained to do so
    - but have important areas of significant ignorance - why?
- Pose profound vision questions
  - maybe regression is over?



Du, ... , Bhattad. CVPR-W 2024

They know about water physics…(?!?)

Vision Language Models

Models using physical inference

Aghajanzadeh et al, in prep

Errors: iffy projective geometry

https://huggingface.co/spaces/stabilityai/stable-diffusion/discussions/1593

Errors: iffy projective geometry

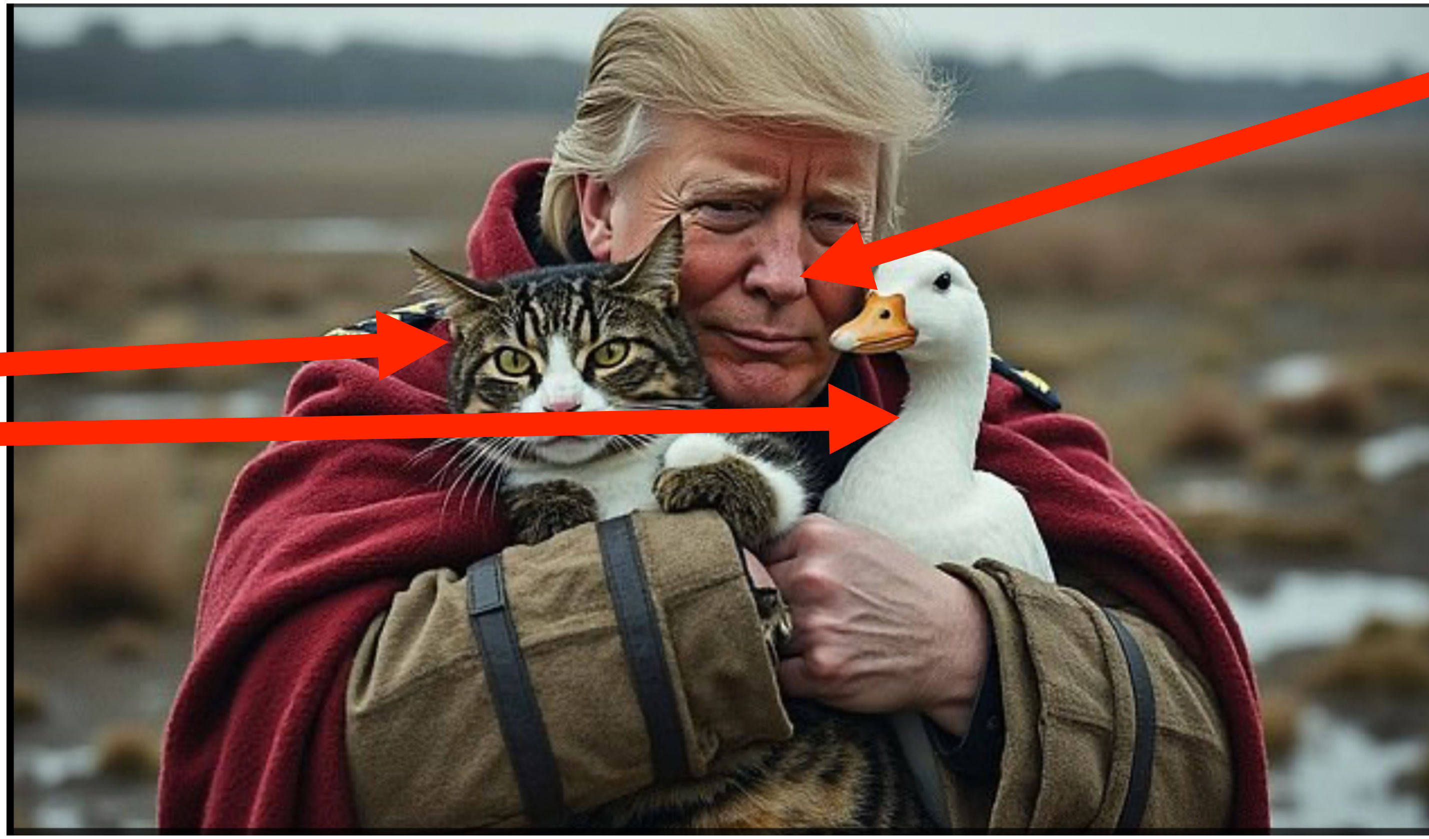
https://huggingface.co/spaces/stabilityai/stable-diffusion/discussions/1593

Errors: iffy projective geometry

https://huggingface.co/spaces/stabilityai/stable-diffusion/discussions/1593

Q: real or fake?

Scale relative to bear



Scale relative to politician

He’s got someone else’s nose!

Do the mistakes matter?

- Yes
  - if you care about understanding things

- No
  - if you're making weird political posters, etc.

This is not the behavior of people who are carefully creating traditional software. This is the behavior of people who are growing a whole new sort of creature and then taking whatever they get.

AI Is Grown, Not Built

(Nobody knows exactly what an AI will become. That's very bad.)

By Eliezer Yudkowsky and Nate Soares, The Atlantic

Main points

- Increasingly large encoders can do increasingly more stuff

- Easy to bodge large encoders together with language models

- Quite a lot of evidence that surprising outcomes emerge
  - Why surprising?
    - Actually, we have very little basis to know what to expect

- This may be pricing itself out of almost any market

Big Picture

- Train models using both images and text
- Models should
  - accept text/images, produce text/images
  - answer questions about images

- Advantages
  - extremely rich training opportunities
    - for a general vision encoder
  - lever text descriptions of objects to build detectors
    - particularly important for rare or unseen objects
      - a version of distributional semantics

An Introduction to Vision-Language Modeling
Bordes et al, 2024

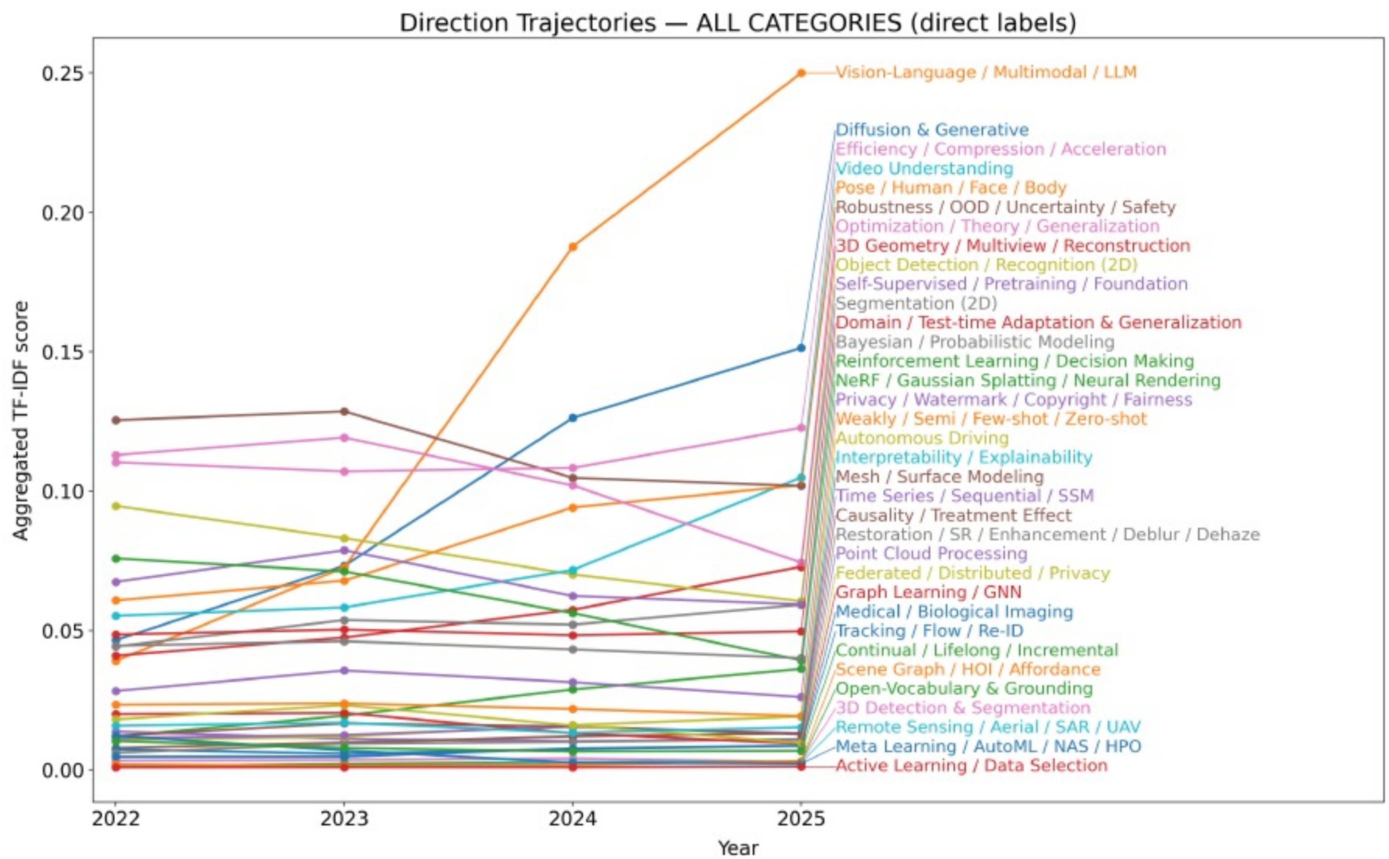


Figure 1: **Direction Trajectories across CVPR+ICLR+NeurIPS — ALL CATEGORIES (direct labels).** Each curve is the yearly aggregated TF–IDF mass for a direction (integer year ticks).

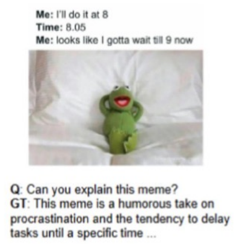
Vision Language Models: A Survey of 26K Papers (CVPR, ICLR, NeurIPS 2023–2025), Lin, 2025

Evaluation strategies - I

1) *Answer Matching*

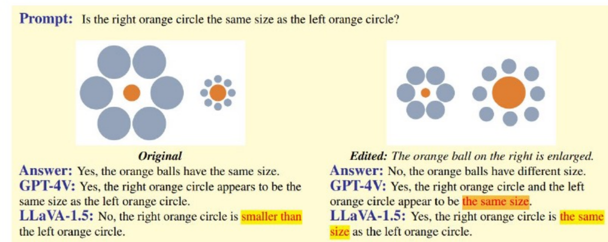
ST-VQA [20]

Evaluation: Accuracy
Metric: Exact Match
Format: Specific short-form answers, such as objects…



MM-Vet [245]

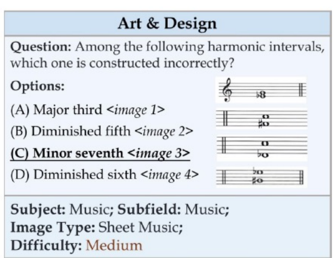
Evaluation: Average Score
Metric: ROUGE, LLM Eval
Format: Long-form open-ended answers



HallusionBench [67]

Evaluation: Accuracy / Precision / Recall
Metric: Exact Match
Format: Yes/No question

2) *Multiple Choice*



MMMU-Pro [248]

Evaluation: Accuracy
Metric: Exact Choice Match
Format: Multiple choice questions

3) *Image-Caption Similarity*

T2I-CompBench [63]

Evaluation: Average Similarity
Metric: CLIPScore, GenEval
Format: text to image generation

A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges, Li et al, 2025

Nuisances: hallucination

TD: A cat is sitting on a bed in a room.
S: 12.1 M: 23.8 C: 69.7
TD Restrict: A bed with a blanket and a pillow on it.
S: 23.5 M: 25.4 C: 52.5

TD: A cat laying on the ground with a frisbee.
S: 8.0 M: 13.1 C: 37.0
TD Restrict: A black and white animal laying on the ground.
S: 7.7 M: 15.9 C: 17.4

Figure 5: Examples of how TopDown (TD) sentences change when we enforce that objects cannot be hallucinated: SPICE (S), Meteor (M), CIDEr (C), see Section 3.4.

- Cure strategy: force entities in captions to have corresponding detector responses

Object Hallucination in Image Captioning, Rohrbach et al 2019

Main points

- Low/efficient compute vision is hugely important
  - and will likely continue to diversify

- Much future action will be here

Very hard open questions

- Transformers are super
    - but ludicrously expensive and hard to train
    - is there something better?
- Big models can do more stuff than little models
    - why?
    - so what?
- Could you learn to produce low-compute models?
- Train a model
    - What do you “expect it to know”?
    - What do you “not expect it to know”?
- Vision routinely breaks “the rules” of machine learning
    - very little bad happens – why?