Beyond Human Language

CSC2541H1 Topics in Machine Learning, Winter 2025, UToronto

Chris J. Maddison

Announcements

Questions?

Today

- Natural language is not the only data in the world.
  - By volume, particle colliders probably produce the largest number of bits*.
- Today's topics
  - Vision Language Models (Flamingo and JetFormer)
  - Genomic Foundation Models (Evo)
- Similar themes emerge, but there are some unique challenges.

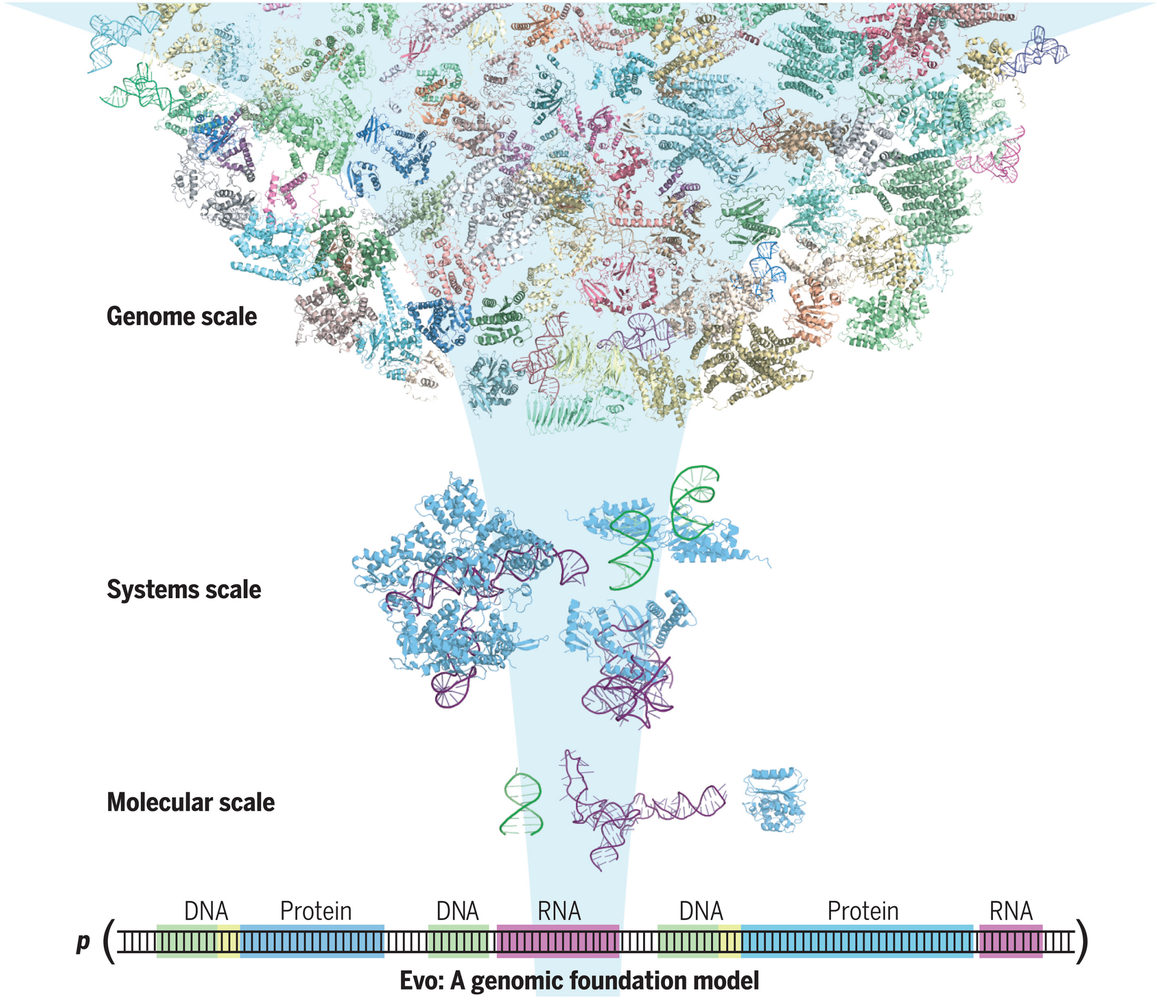
* Clissa et al. 2023. How big is Big Data? A comprehensive survey of data production, storage, and streaming in science and industry.

Long Contexts

Challenge 1

- Scale along the time axis is a serious challenge in some domains (video, genomics)
- Genomes are sequences of nucleotides that store the instructions for the synthesis of biomolecules in cells.
- Genomes have small alphabets (A, C, G, T for DNA) but are very long, up to 160B nucleotides*.

* Fernández et al. 2024. A 160 Gbp fork fern genome shatters size record for eukaryote.

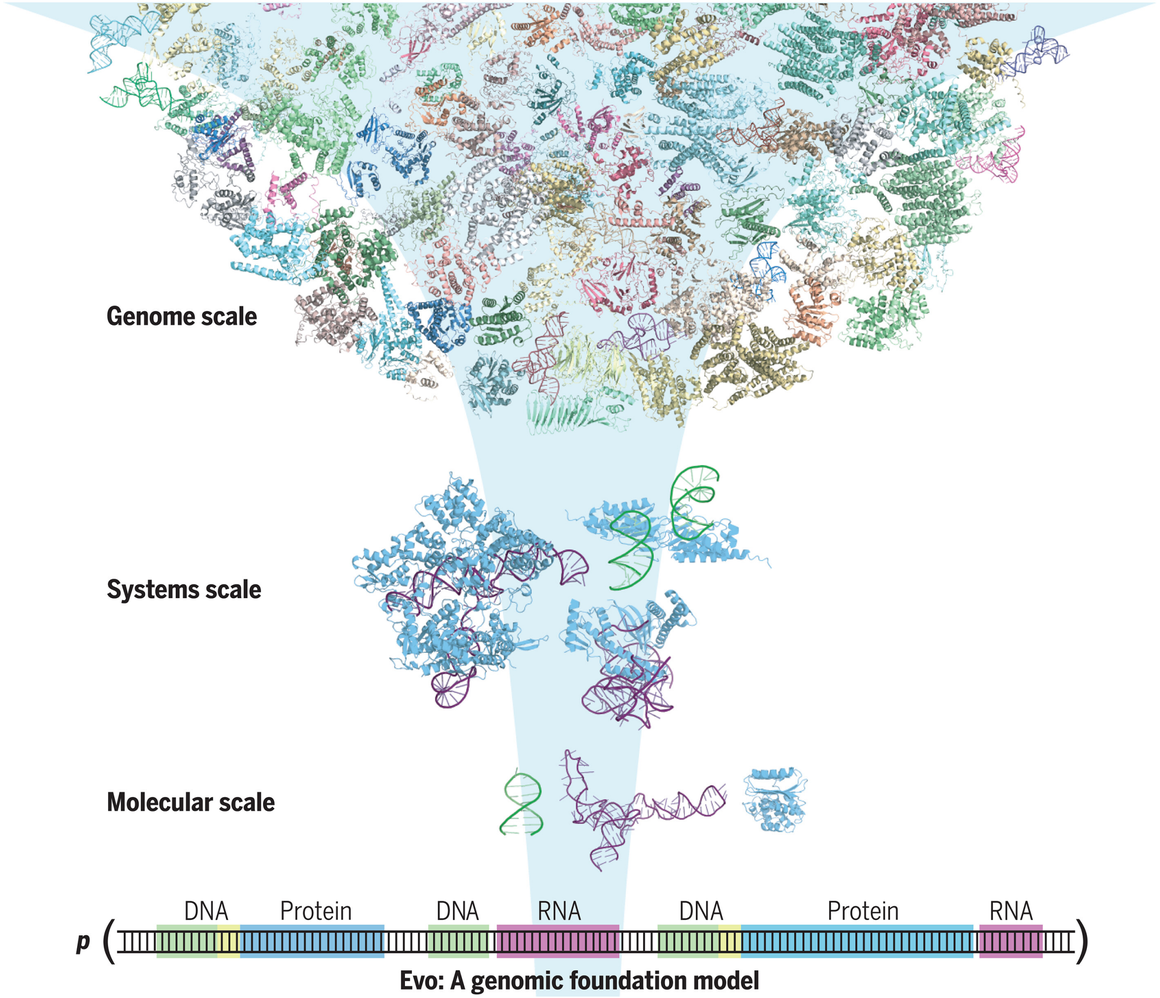


Nguyen et al. 2024. Sequence modeling and design from molecular to genome scale with Evo.

Long Contexts

Challenge 1

- Arguably a genome is similar to a document or a novel or a software library.
- Would love to fit a whole genome in context, but this is challenging if we model at a single-nucleotide resolution.

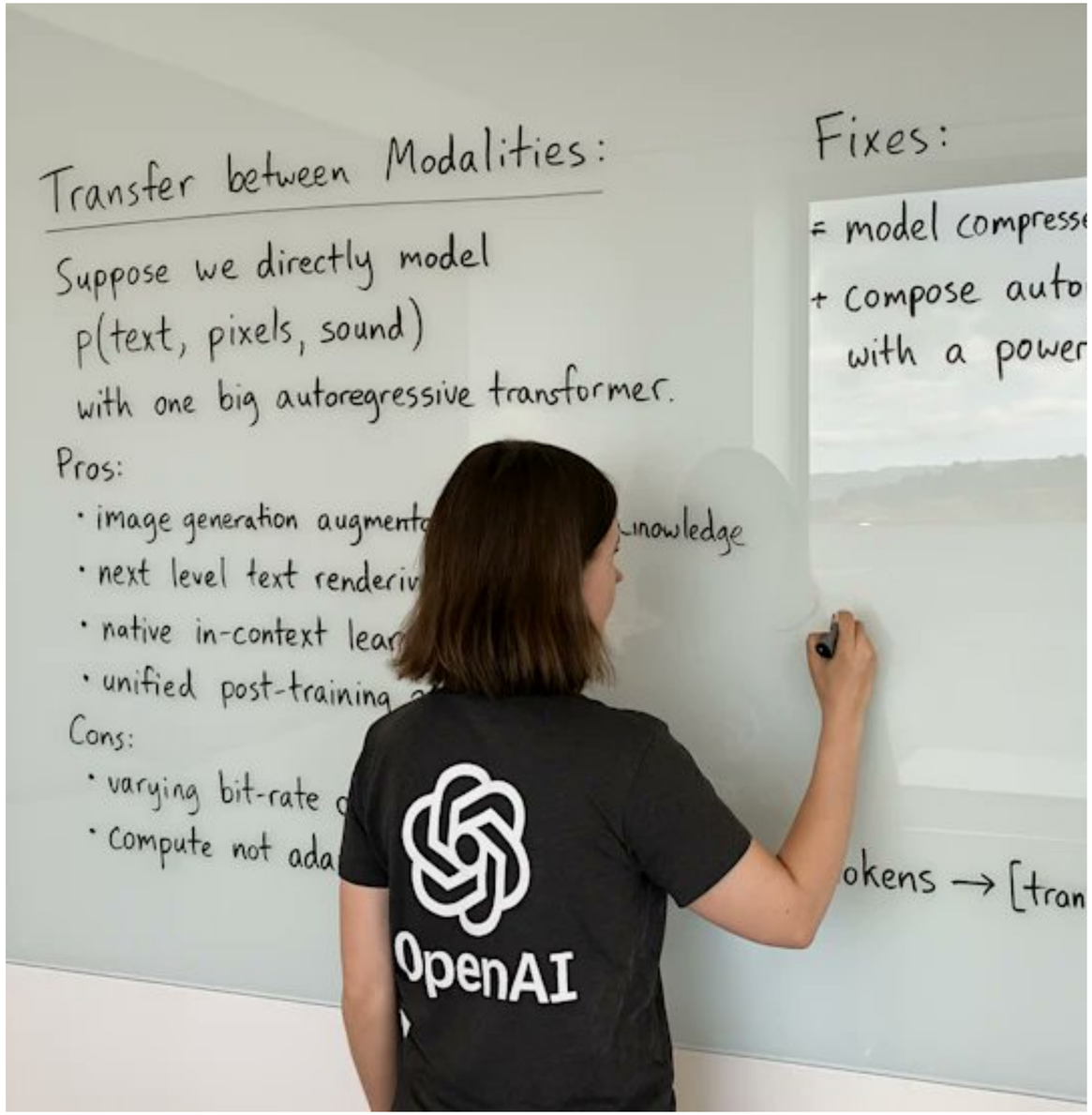


Nguyen et al. 2024. Sequence modeling and design from molecular to genome scale with Evo.

Multimodality

Why?

- **Multimodality tends to refer to combining data types** from different measurements (e.g., image and text).
- Some data is not modelled well by sequences of a finite alphabet.
  - If it has geometric structure, e.g., weather on the surface of the earth, that we want to model.
- Humans prefer to communicate multi-modally.
  - Even if it's inefficient from an information theoretic perspective, we sometimes prefer to communicate visually.

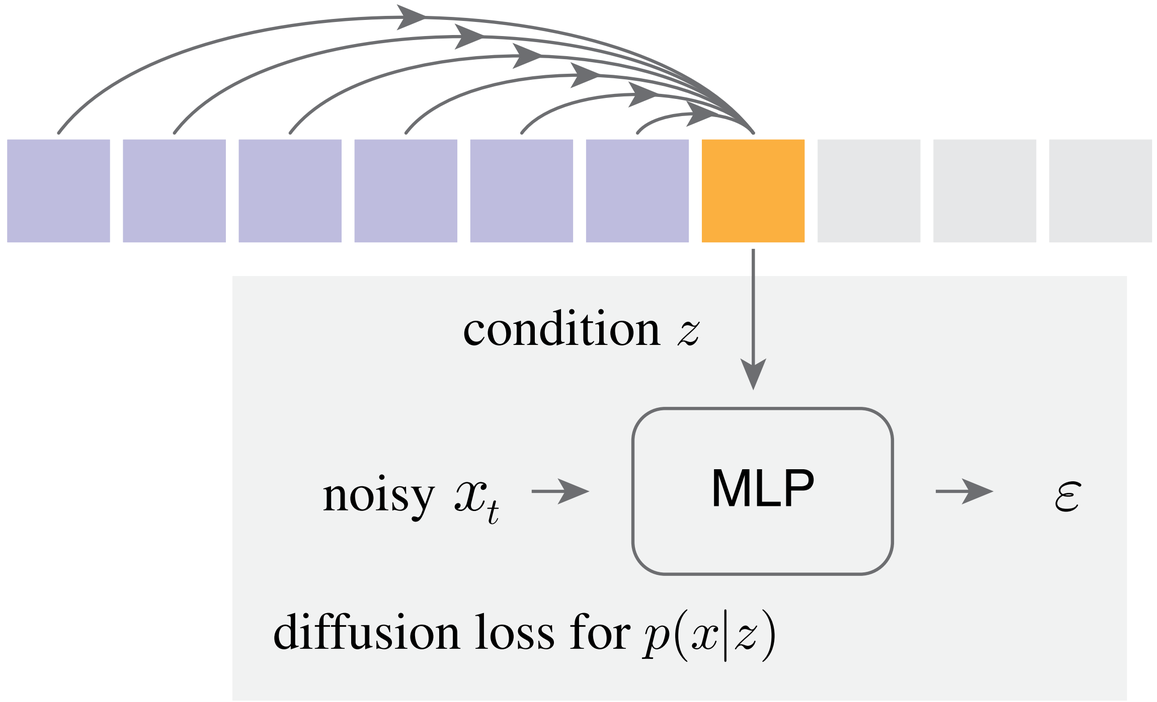


Synthetic image from GPT 4o native image generation release.

Multimodality

Challenge 2

- **Ideally: a large model trained on the joint over all data types.**
- Possible in principle:
  - Transformers produce embeddings of sequences to parameterize a distribution over a finite alphabet.
  - We can use the next token embedding to parameterize a distribution over any space.

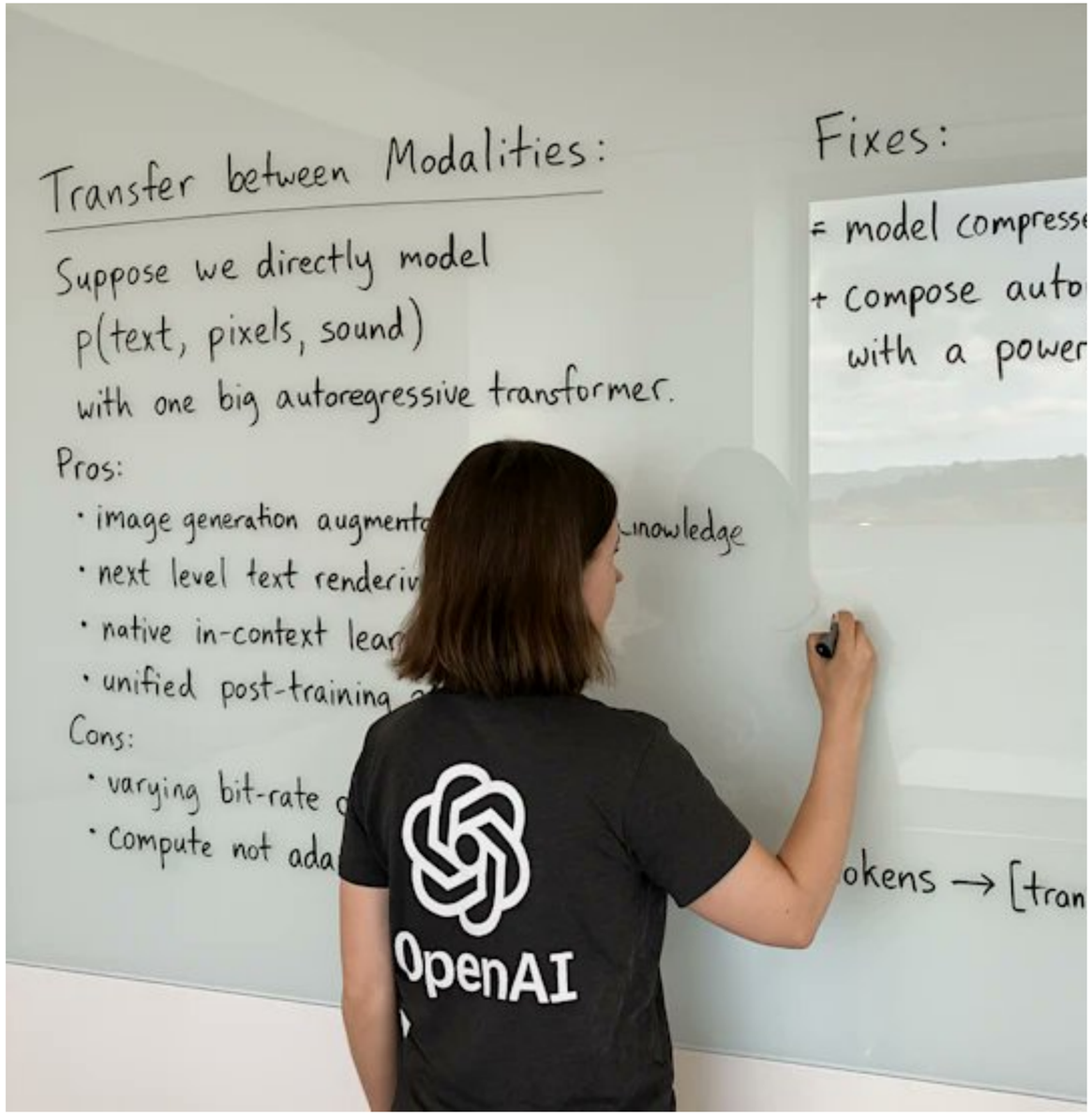


Tianhong et al, 2024, Autoregressive Image Generation without Vector Quantization

Multimodality

Challenge 2

- Key challenge: loss magnitudes can vary wildly across modalities.
- One way to think about this is to think about compression.
- The more information contained in a modality, generally the more it will dominate the loss.



Synthetic image from GPT 4o native image generation release.

Multimodality

Challenge 2

- How much information is in different modalities?
- **Finite alphabets:** at most log2(N) bits to store where N is the size of the alphabet.
- **Real-value alphabets:** You cannot store a real number with infinite precision any any number of bits.
- This is why we have lossy compressors like JPEG for images.

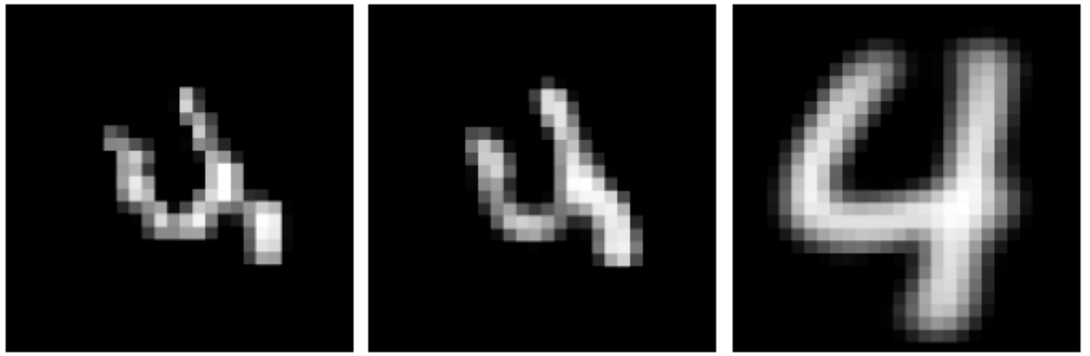
High bitrate

Low bitrate

Multimodality

Challenge 2

- Lossy compressors have to decide what to model.
  - Ideally we model “what matters to humans”
  - But this is not well-specified.
- In other words, what we choose to model in the loss is not well-specified and has an impact on the bitrate of that modality.

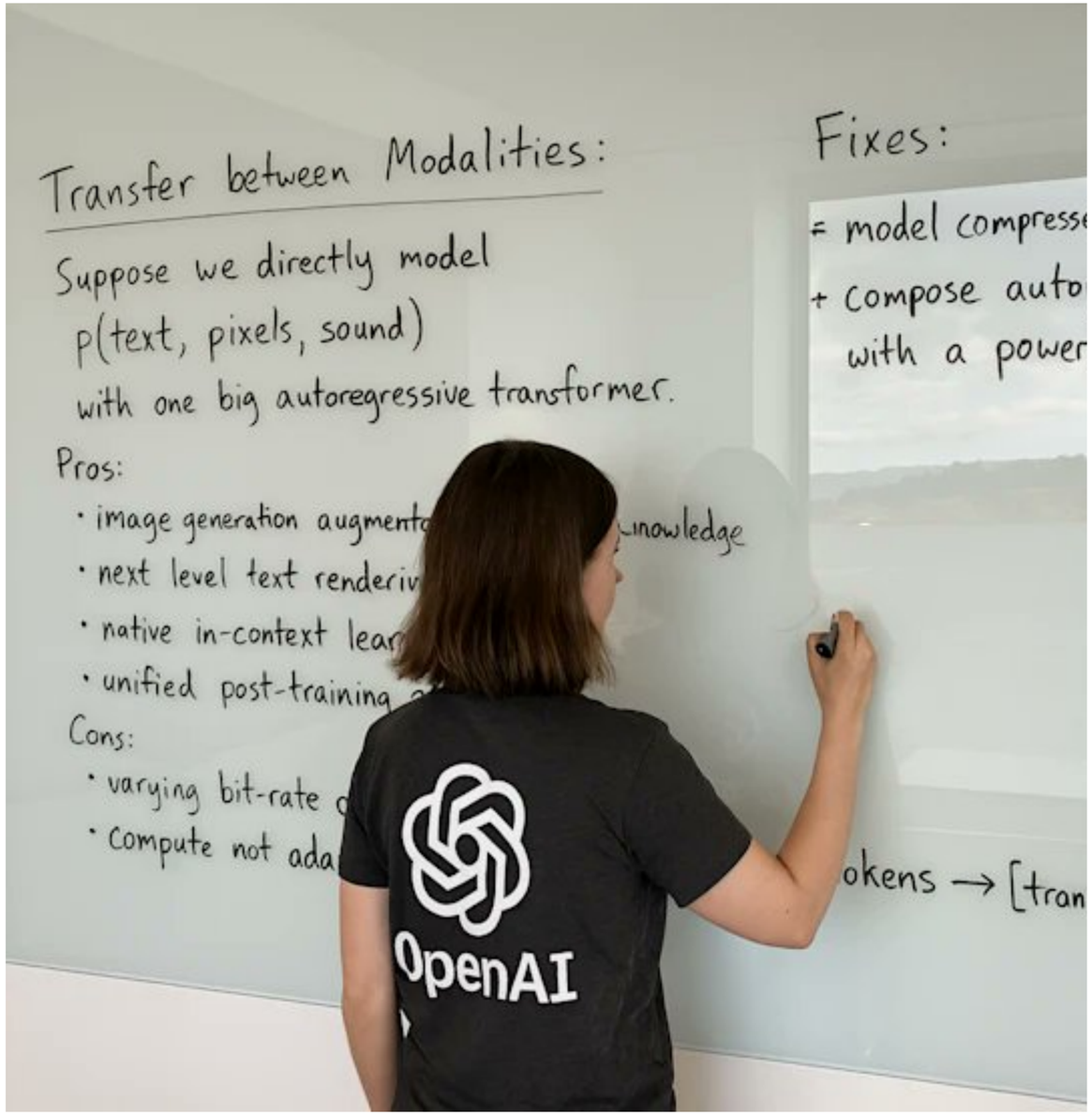


Source Standard rec. Our rec.

Multimodality

Challenge 2

- Varying bitrate between modalities can hurt model training because sources with very high bitrate are overweighted in the loss.
- Take-home: high bitrate modalities dominate naive training losses like cross-entropy but do not reflect what humans necessarily care about.



Synthetic image from GPT 4o native image generation release.