Weaviate

Scaling MultiModal Models with Vector Databases

Zain Hasan

Haystack EU 2023 – 21st, September 2023

Who here thinks: AI poses an existential threat to humans in the next 5 years?

Yes


BE IT RESOLVED,
AI research and development poses an existential threat.
The Munk Debate • Toronto • June 22, 2023
"BE IT RESOLVED: AI research and development poses an existential threat."
67%
33%
✓ PRO
✓ PRO
✕ CON
✕ CON
VS
Yoshua Bengio
Max Tegmark
Melanie Mitchell
Yann LeCun

Why don’t we have AI that can:

- Drive
- Cook a meal
- Setup the table
- Walk naturally
- … you name it!

Moravec's Paradox

For AI:

Mind-numbingly easy

- Language translation
- Playing chess
- Calculus

For Humans:

1508

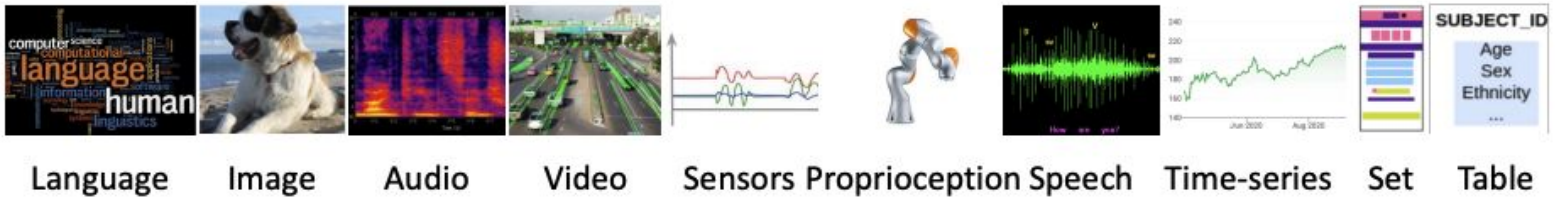
It's the opposite! ... WHY!?

How do humans learn?

… a lot of learning is multimodal & non-lingual!

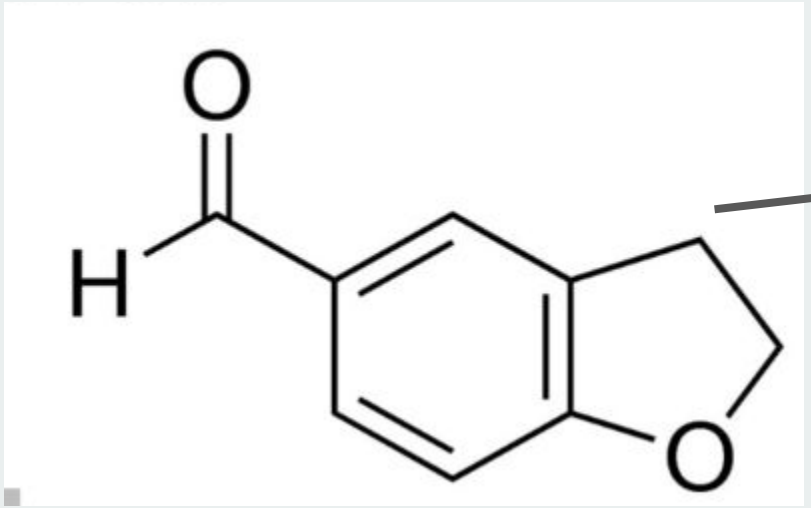
The Promise of Multimodal Models!

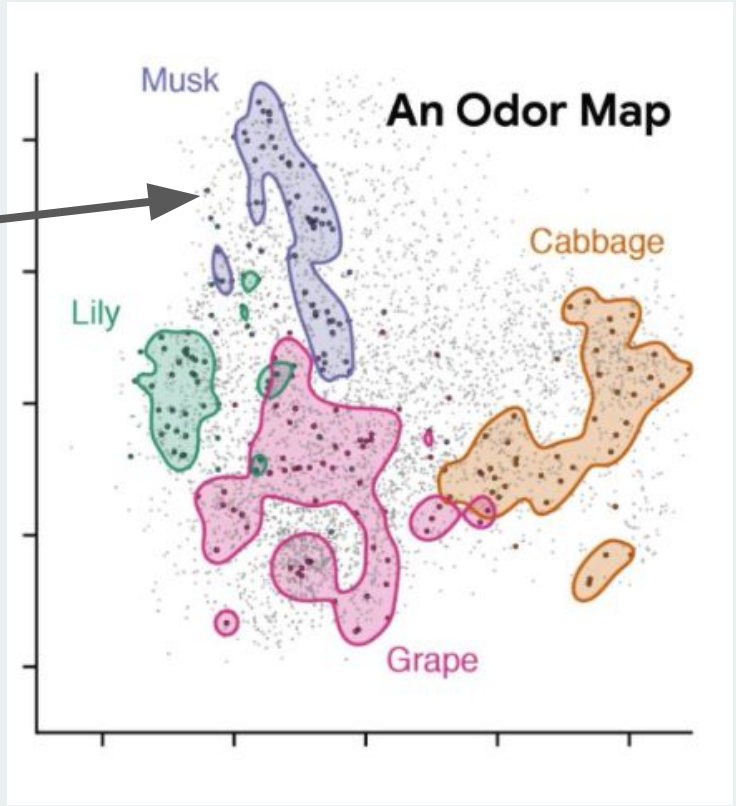
Understand a datapoint from multiple modalities



Source: High-Modality Multimodal Transformer- https://arxiv.org/pdf/2203.01311.pdf

You can even digitize smell!

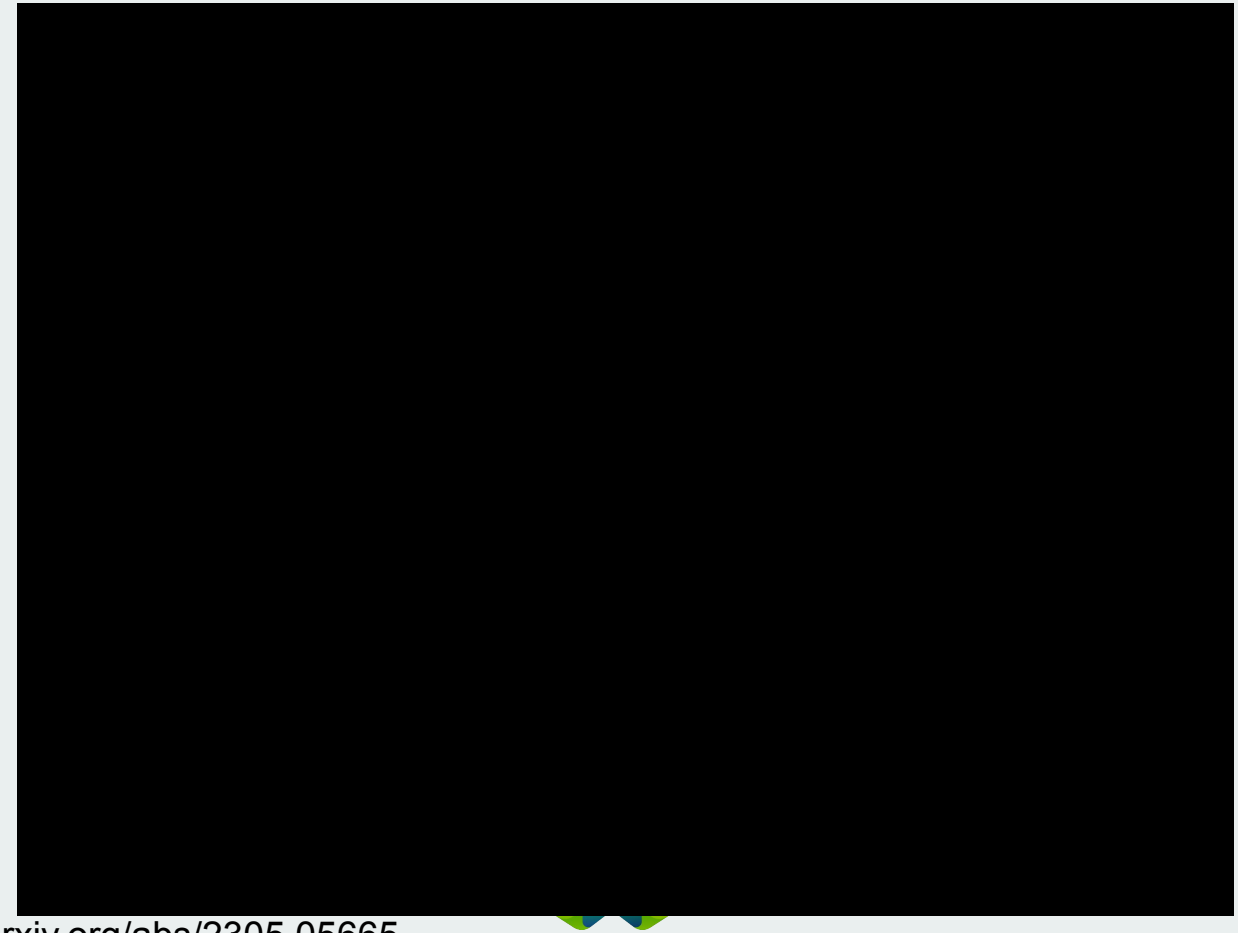




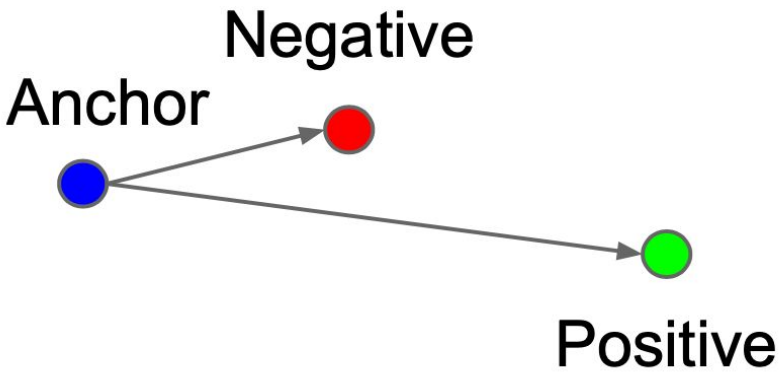
https://blog.research.google/2022/09/digitizing-smell-using-molecular-maps.html
https://arxiv.org/pdf/1910.10685.pdf

How do Multimodal Models work?

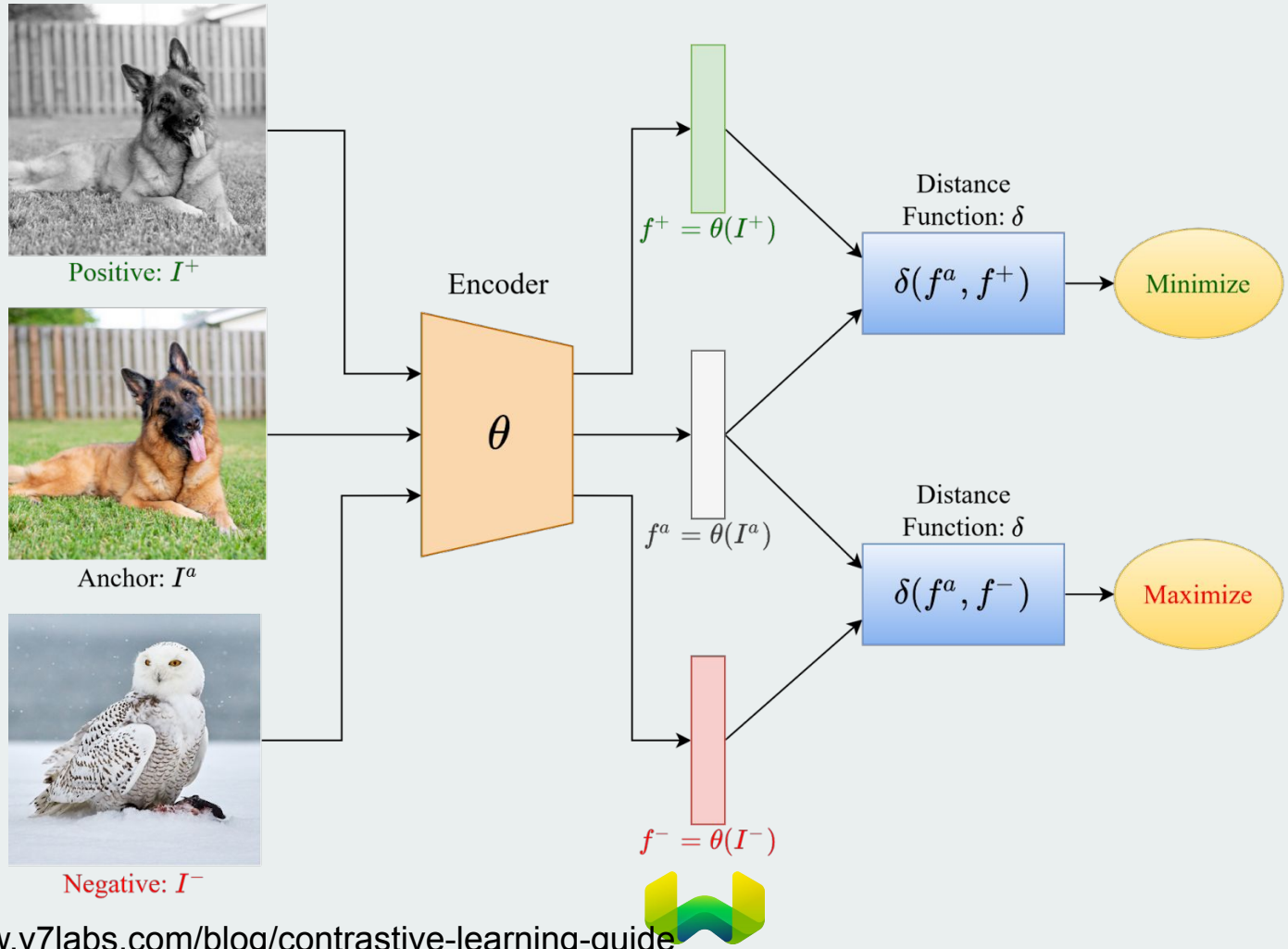
Train one model per data type and then unify them!



ImageBind: https://arxiv.org/abs/2305.05665

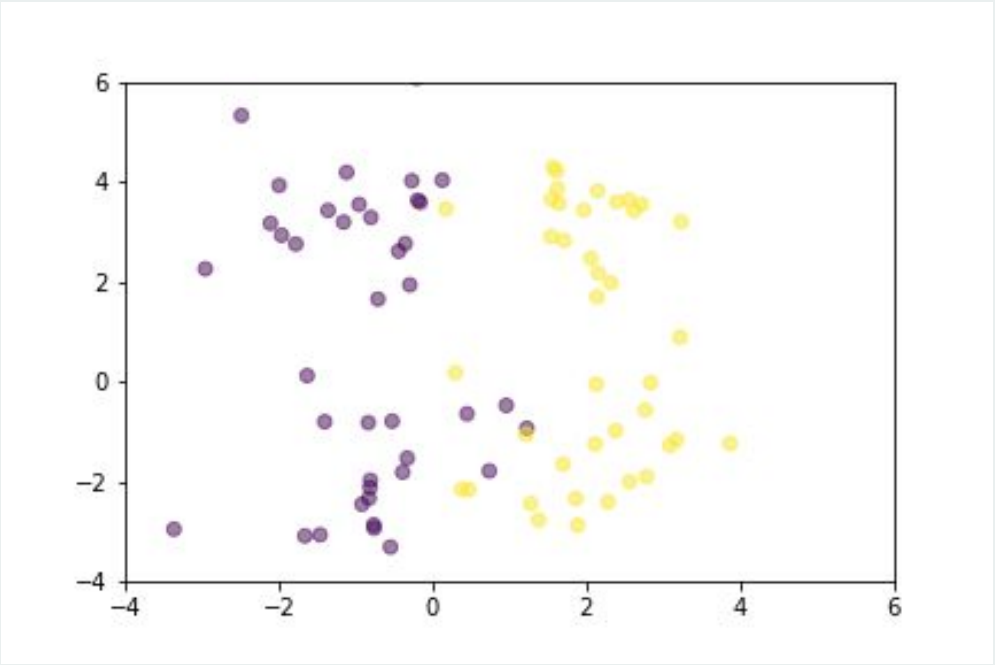


https://lilianweng.github.io/posts/2021-05-31-contrastive/



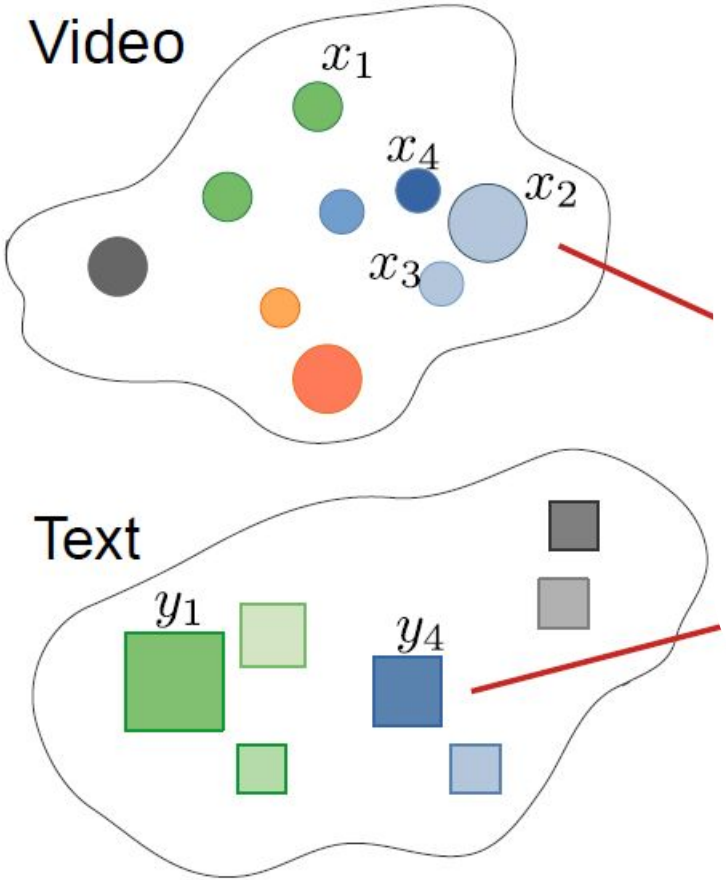
https://www.v7labs.com/blog/contrastive-learning-guide

In action this looks like:



https://medium.com/@maksym.bekuzarov/losses-explained-contrastive-loss-f8f57fe32246

Cross-Modal Contrastive Learning



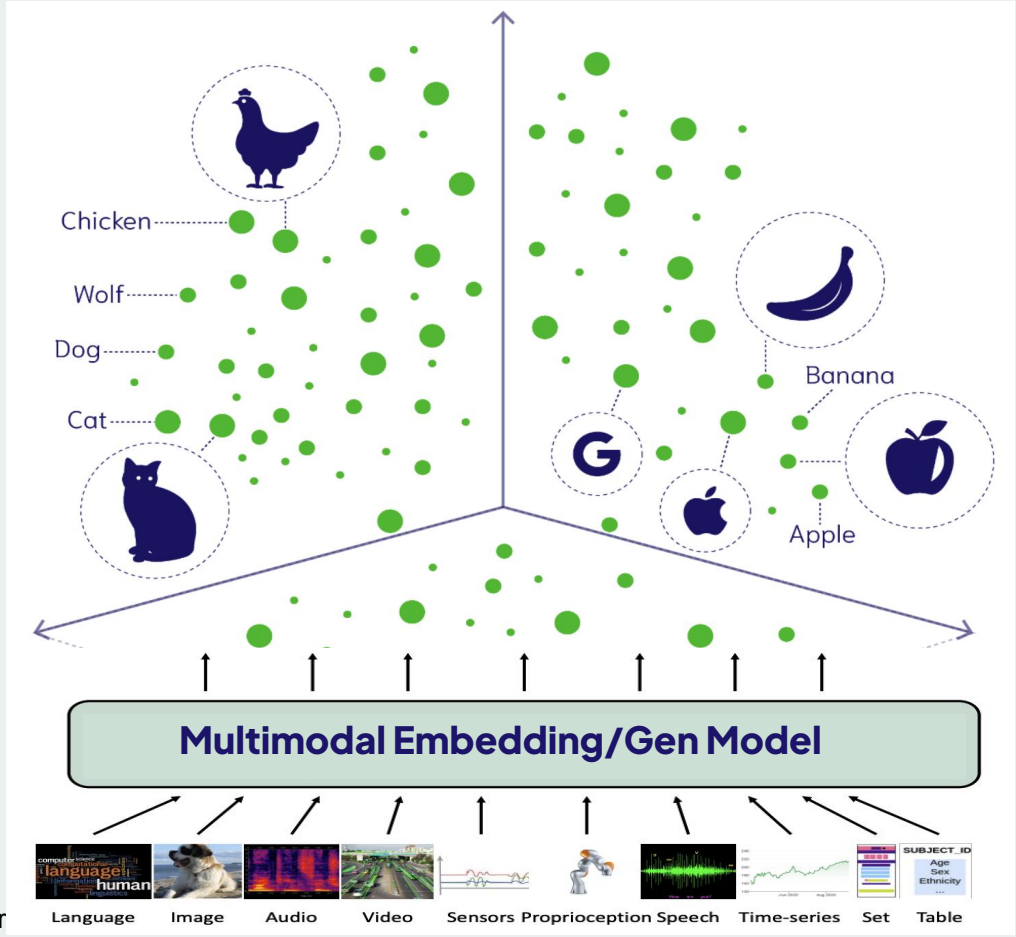
https://lmb.informatik.uni-freiburg.de/Publications/2021/ZB21/
Contrastive Representation Learning: A Framework and Review - https://arxiv.org/abs/2010.05113

InfoNCE Loss Function allows us to do this unification:

$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}, \quad (1)$$

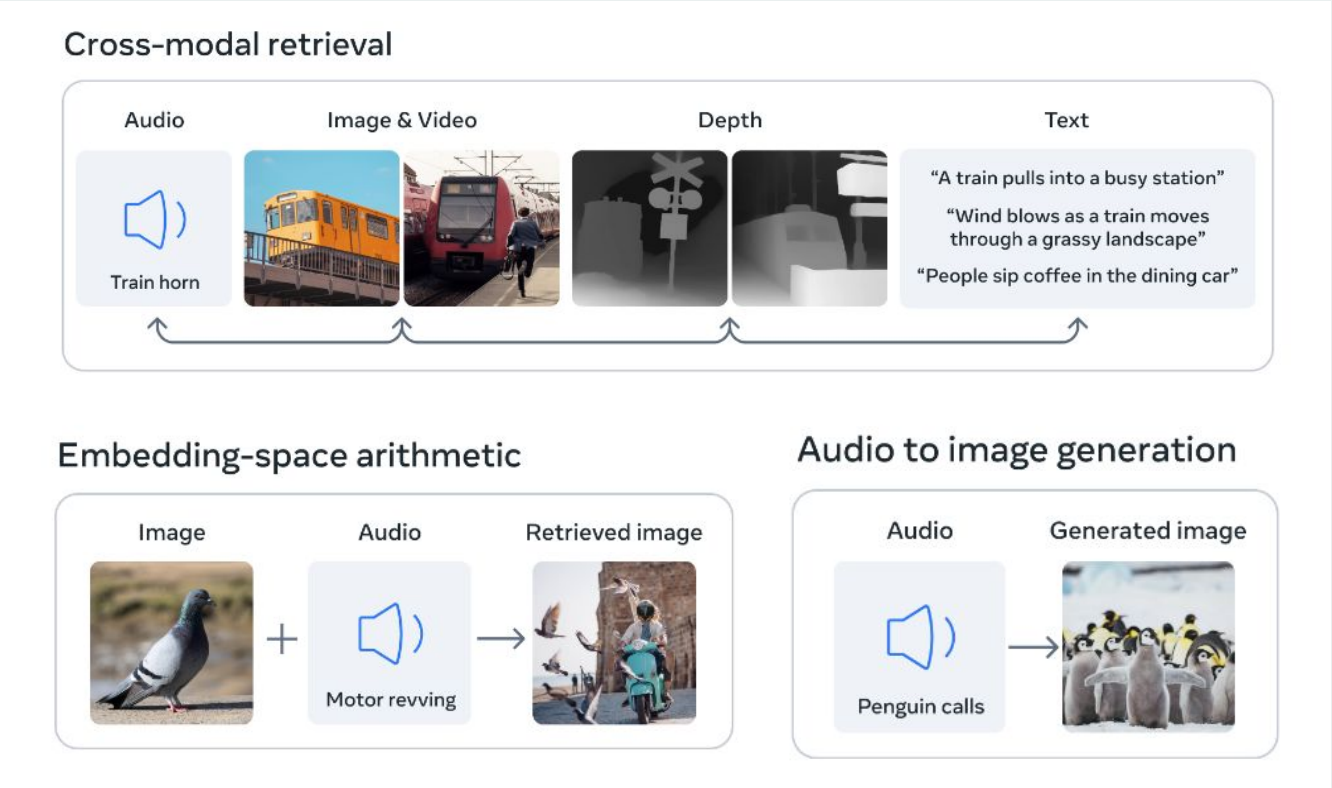
$$\mathbf{q}_i = f(\mathbf{I}_i) \text{ and } \mathbf{k}_i = g(\mathbf{M}_i)$$

This generates one unified vector space



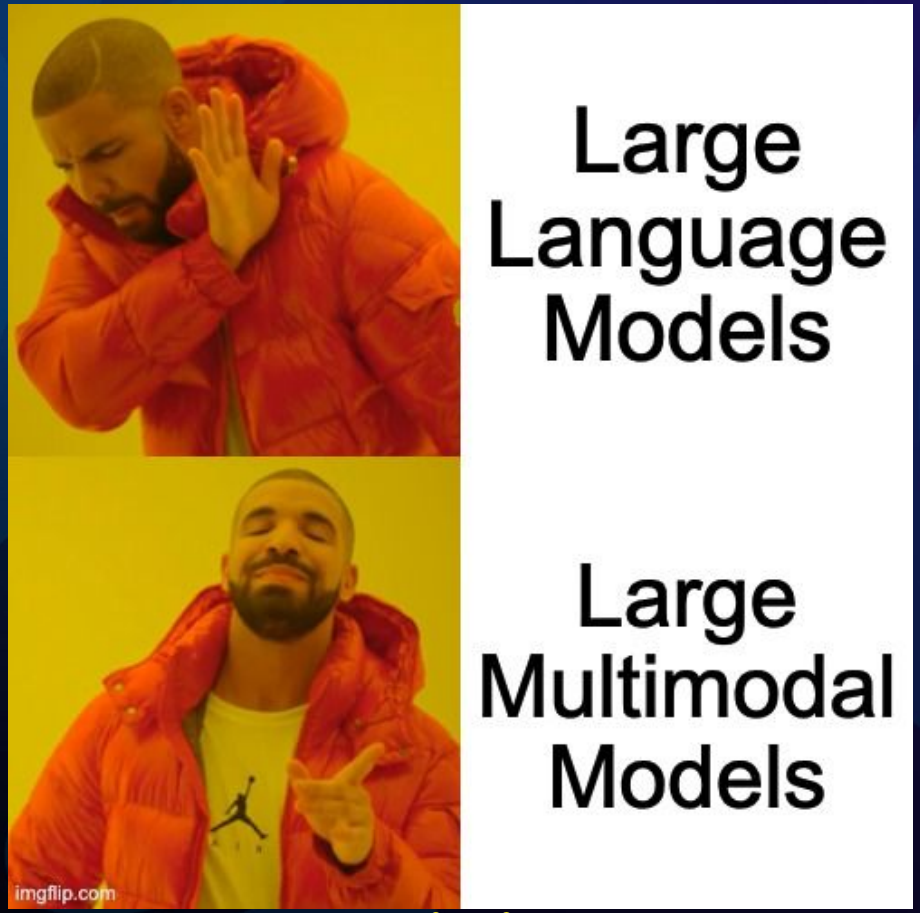
Source: High-Modality Multimodal Tr

Cross Modal Functionality …Reasoning



Source: ImageBind: One Embedding Space To Bind Them All - https://arxiv.org/abs/2305.05665

Demo!

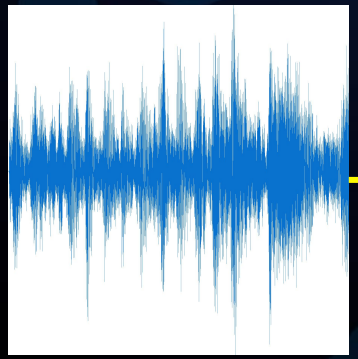

Large
Language
Models
Large
Multimodal
Models
imgflip.com

Applications of MM Models + Vector Databases

Improving E-Commerce Recommender Systems

What burger do you like?

"Juicy, big, loaded with toppings of my choice. High quality beef medium to well with cheese and bacon on a multigrain bun. A huge single or triple burger with all the fixings, cheese, lettuce, tomato, onions and special sauce or mayonnaise!"



Nutrition Facts
Valeur nutritive
Per 1 burger (75 g) / pour 1 galette (75 g)

Amount / Teneur	% Daily Value / % valeur quotidienne
Calories / Calories 100	
Fat / Lipides 2 g	**3** %
Saturated / saturés 0.2 g + Trans / trans 0 g	**1** %
Cholesterol / Cholestérol 0 mg	
Sodium / Sodium 320 mg	**13** %
Potassium / Potassium 400 mg	**11** %
Carbohydrate / Glucides 7 g	**2** %
Fibre / Fibres 2 g	**8** %
Sugars / Sucres 1 g	
Protein / Protéines 13 g	

[0.65, 0.15, ..., 0.23, 0.75]

[0.23, 0.45, ..., 0.84, 0.23]

[0.03, 0.97, ..., 0.27, 0.26]

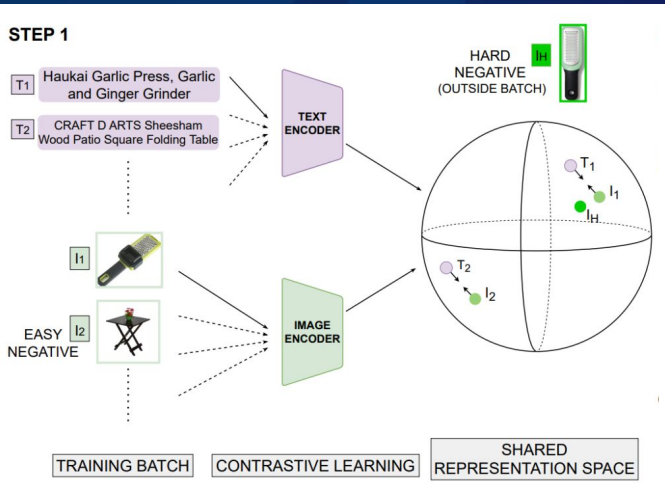
For RecSys MM representations allow us to:

- More uniquely identify what customers like
  - With MM we have more “senses”/modalities to do so
- More uniquely compare relevance b/w products
  - Compare across modalities
- Identify differences amongst similar products

(b) Query Product

Source: https://www.amazon.science/publications/unsupervised-multi-modal-representation-learning-for-high-quality-retrieval-of-similar-products-at-e-commerce-scale

Que2Engage: Embedding-based Retrieval for Relevant and Engaging Products at Facebook Marketplace - https://arxiv.org/pdf/2302.11052.pdf



Source: https://www.amazon.science/publications/unsupervised-multi-modal-representation-learning-for-high-quality-retrieval-of-similar-products-at-e-commerce-scale

We can do even better … some points to consider:

- Why combine modalities by equal weighting?
  - Not all modalities are equal for RecSys tasks
- Why stick for unimodal queries?
  - Multimodal queries can specify details better

Learning Modality Weightings/Importances Using SHAP

- Shapley values allow us to measure individual contributions to a model outcome

SHAP_Taste = 2.25
SHAP_Look = 1.21
SHAP_Text = 1.67

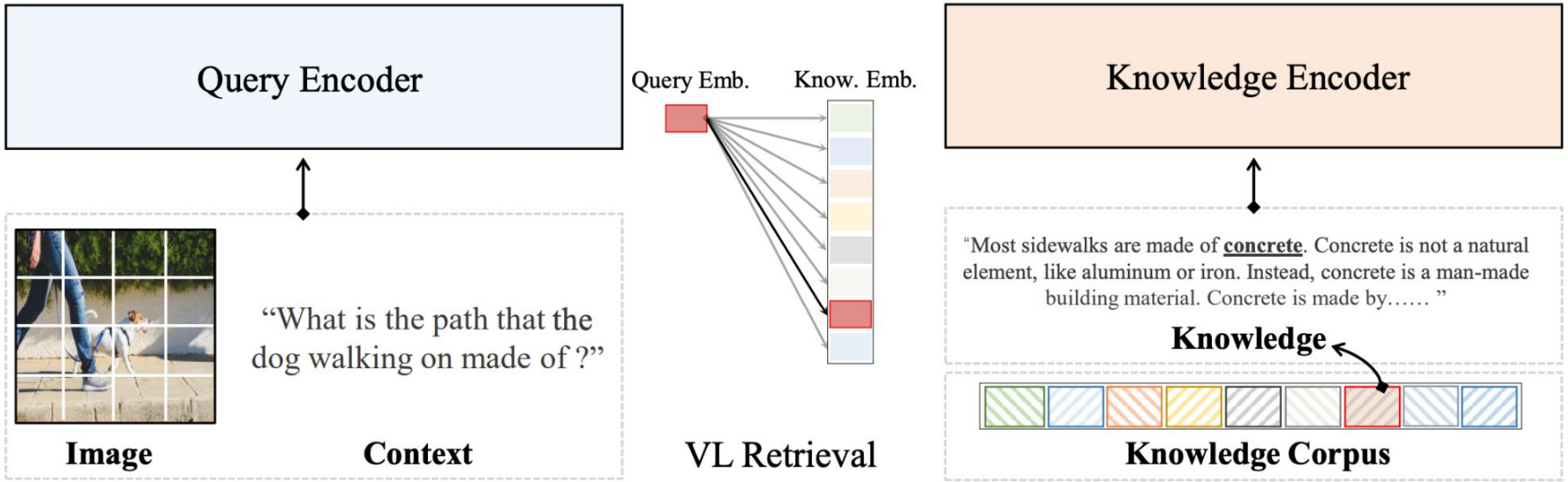
SHAP_Taste = 0.02
SHAP_Look = 3.21
SHAP_Text = 1.43

- Compare an MM models reliance on different modalities
- How much a given modality matters for a given task and dataset

MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks: https://arxiv.org/pdf/2212.08158.pdf

Building Image + Text Queries

- Different techniques to craft text + image queries



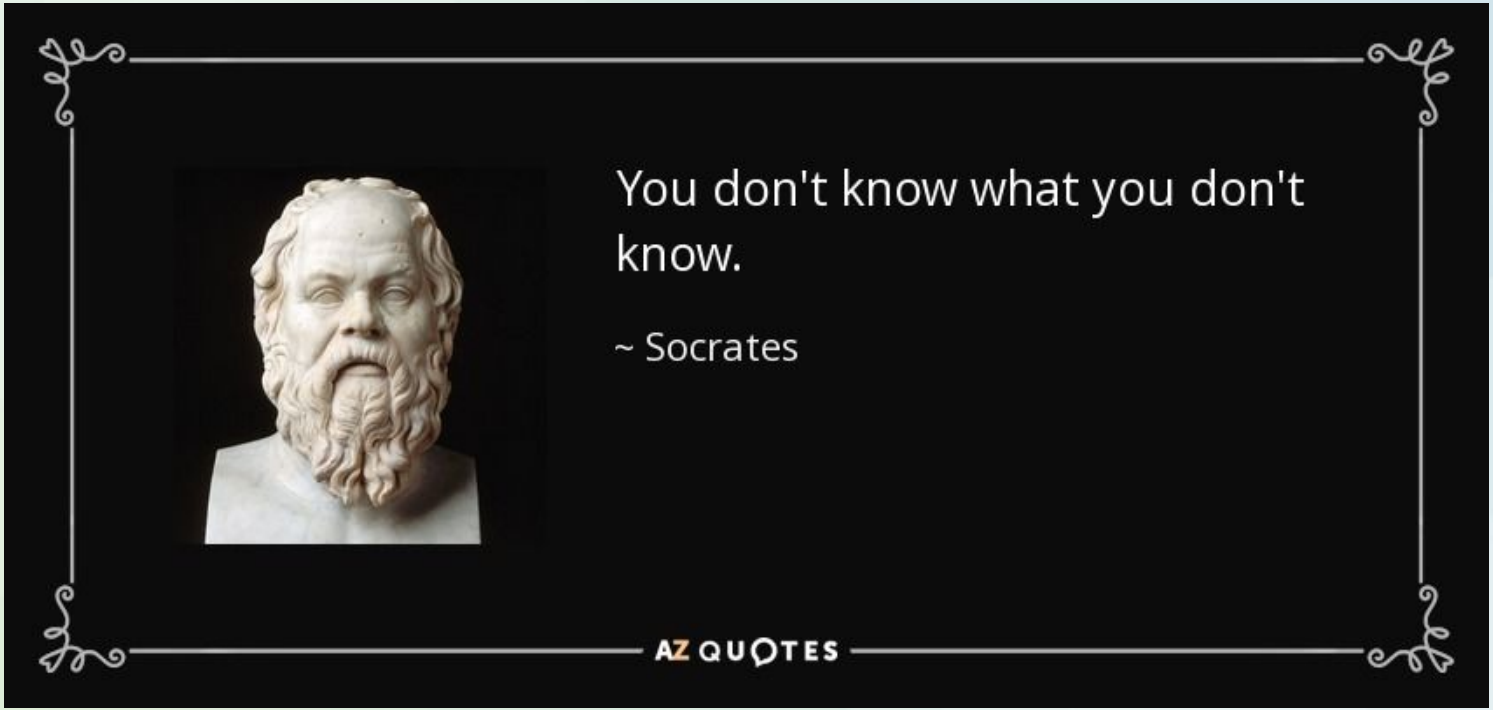
End-to-end Knowledge Retrieval with Multi-modal Queries: https://arxiv.org/pdf/2306.00424.pdf

Building Image + Text Queries

https://openaccess.thecvf.com/content/ICCV2021/papers/Changpinyo_Telling_the_What_While_Pointing_to_the_Where_Multimodal_Queries_ICCV_2021_paper.pdf

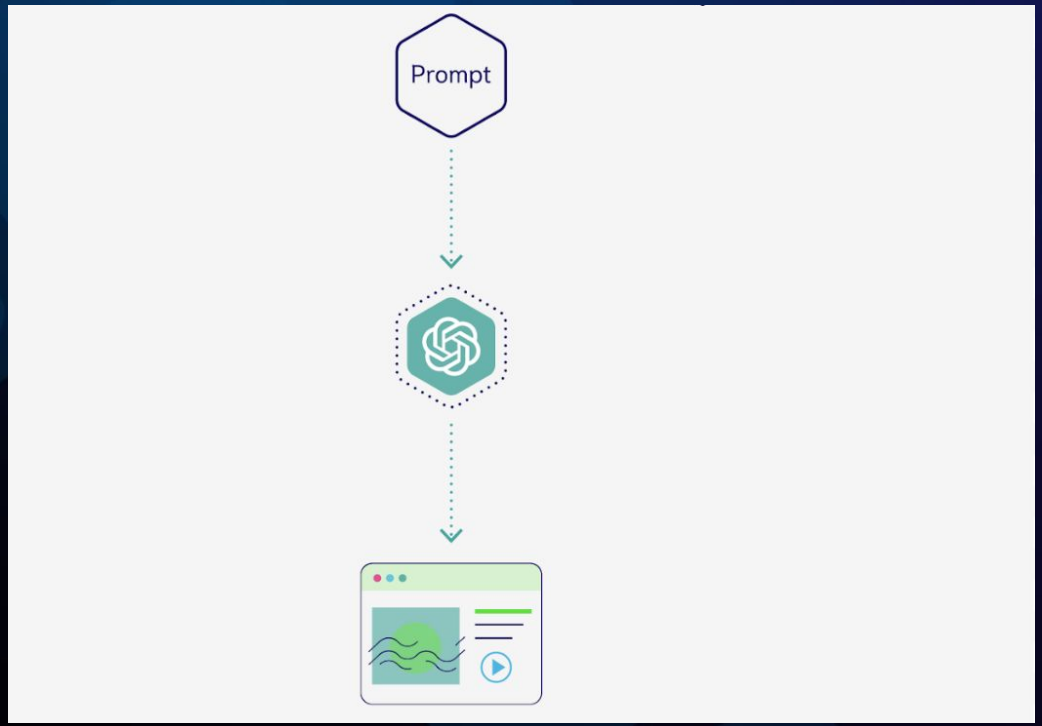
MultiModal Retrieval Augmented Generation (MM-RAG)

A problem with Generative Models



Potential Solution: Answer my prompt ...

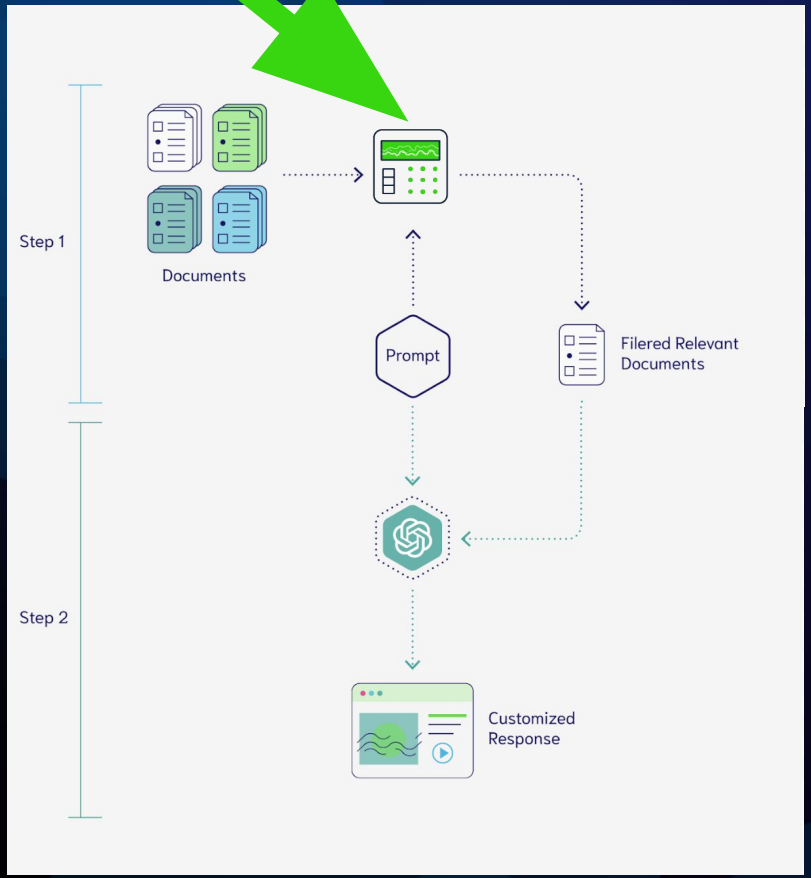
... here's everything relevant you need to know.



This is how the Retrieval Augmented Gen. works!

Visit a vector DB and use vector search to retrieve source material

… then generate answer.

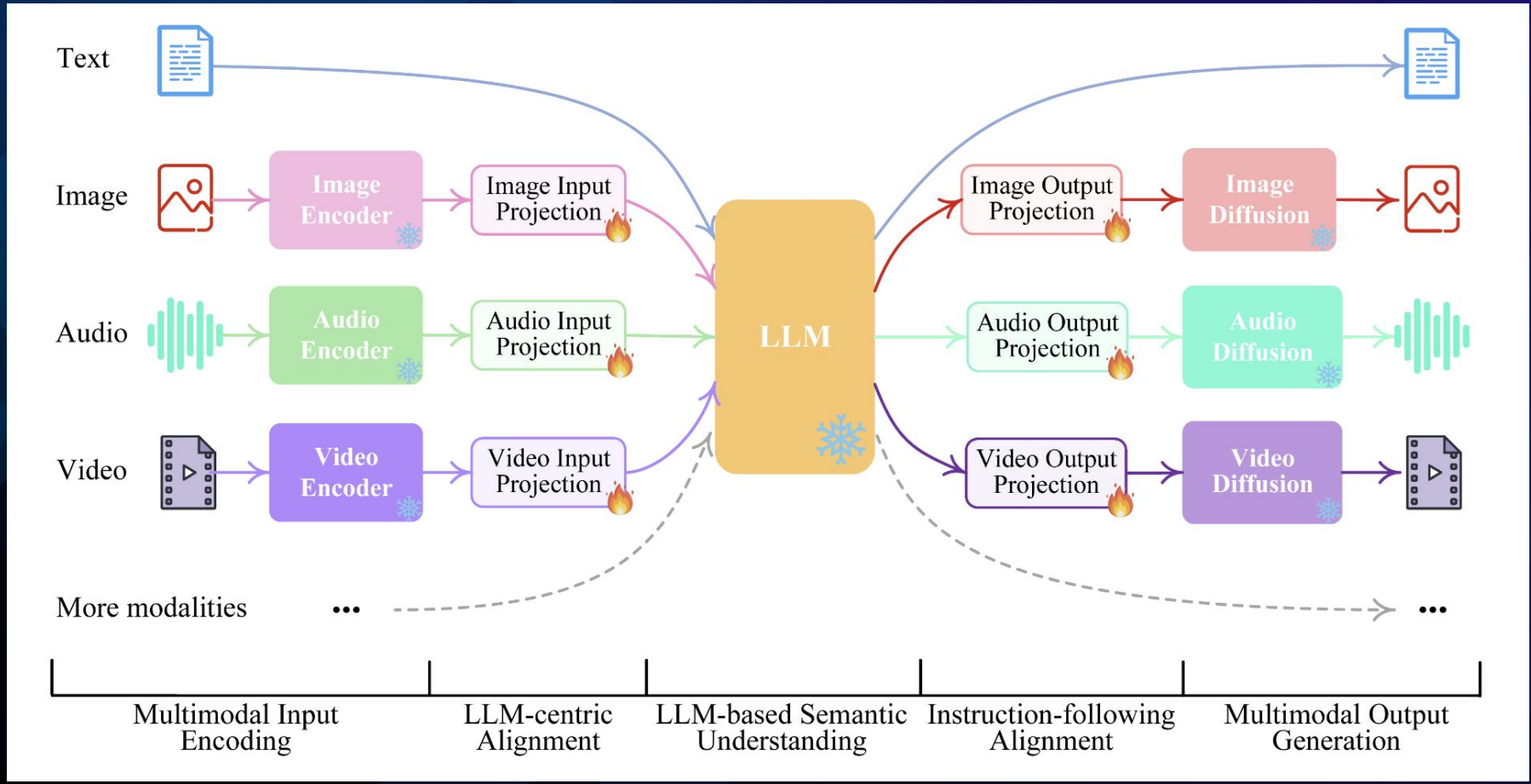


Search over billions of documents in milliseconds.

Multimodal Retrieval and Generation

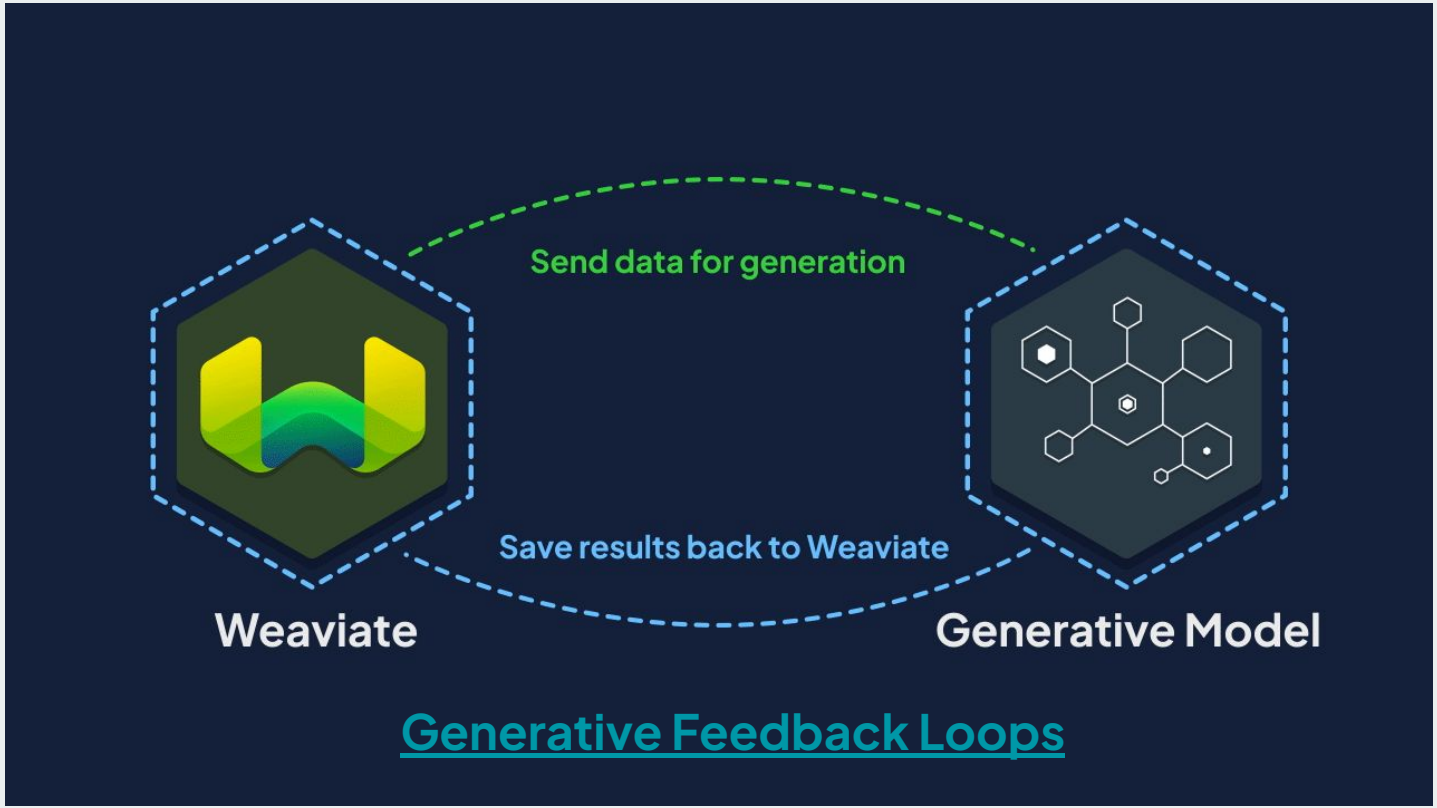
Labrador sitting on bench
near water.

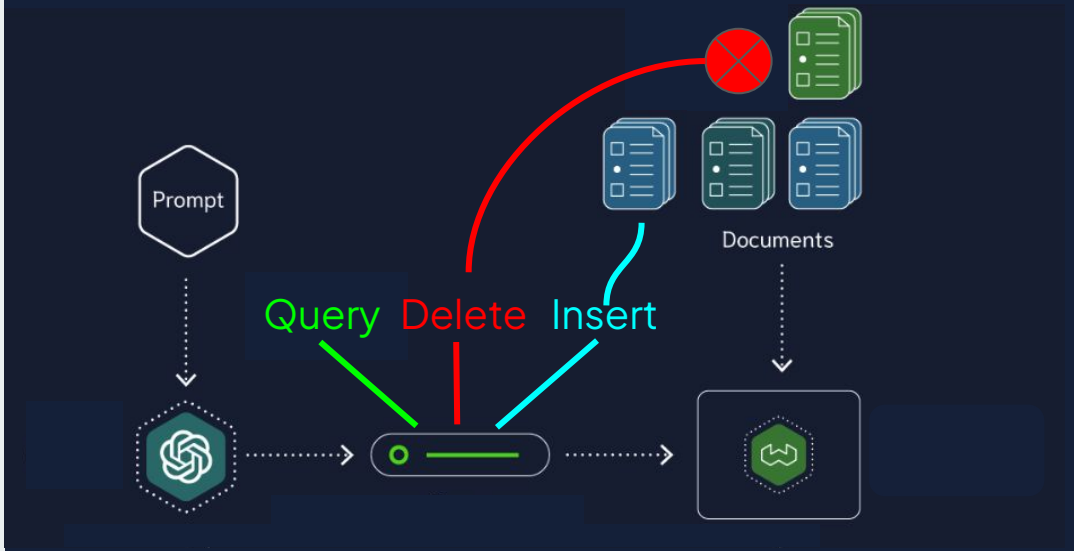
Source: Retrieval-Augmented Multimodal Language Modeling: https://arxiv.org/abs/2211.12561



NExT-GPT:Any-to-Any Multimodal Large Language Model: https://arxiv.org/pdf/2309.05519.pdf

More than just retrieve and generate … you can get MM Gen Models to remember and forget as well!





- Retrieve Data (search query)
- Forget memory (delete objects)
- Remember interaction (insert objects)

MM Search is the Future!

Thank you!


weaviate.io


weaviate/weaviate


@weaviate_io

Reach out!

Zain Hasan


@zainhasan6


linkedin.com/in/zainhas/


@zainhsn