Retrieval-augmented Multimodal Foundation Models

Michihiro Yasunaga

Stanford University

Center for Research on Foundation Models

Stanford University

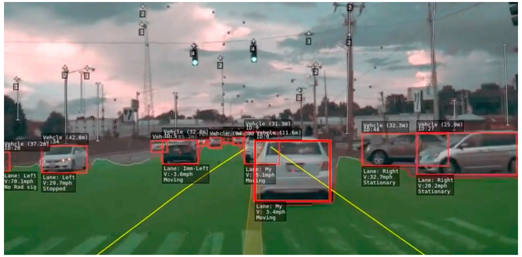
AI is becoming multimodal

Personal Assistants

Search

Generative AI

Autopilot



Multimodal Foundation Models (Text-to-Image)

DALL·E, **Parti** (text ⟶ image; Transformer)

DALL·E 2, **StableDiffusion** (text ⟶ image; Diffusion)

TEXT PROMPT

an armchair in the shape of an avocado. . . .

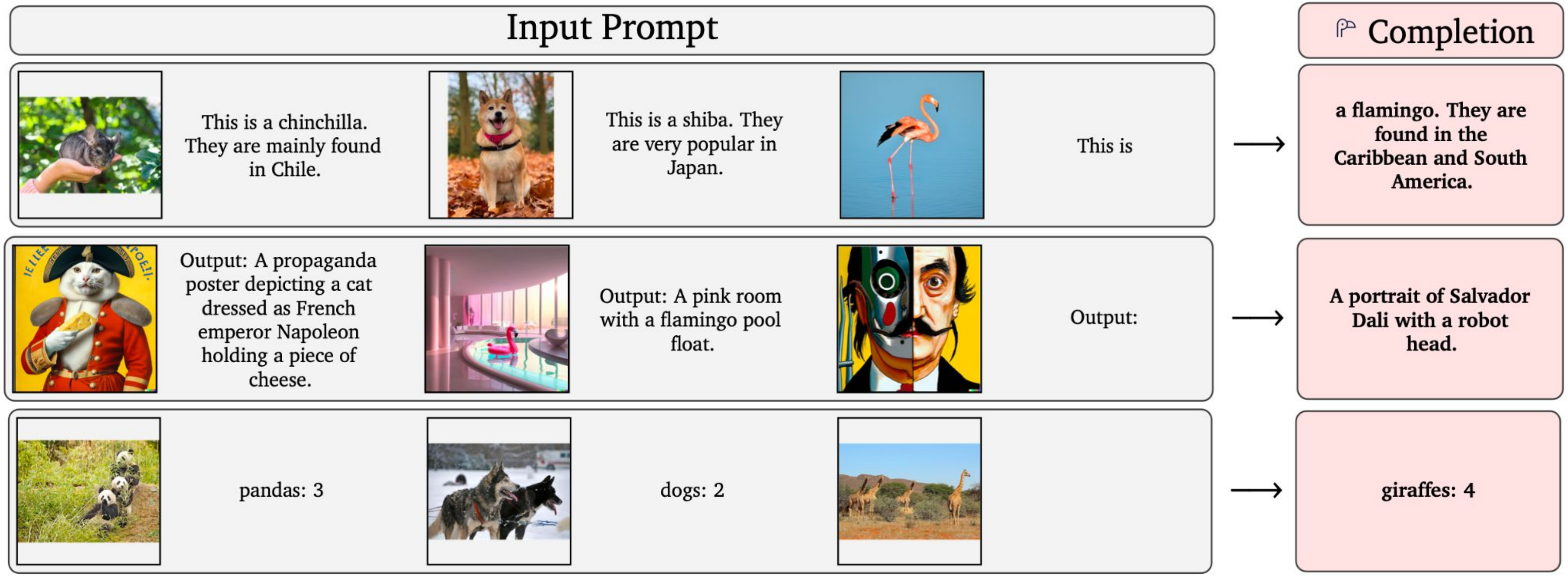
AI-GENERATED IMAGES

Parti-350M | Parti-750M | Parti-3B | Parti-20B

A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!

Multimodal Foundation Models (Image-to-Text)

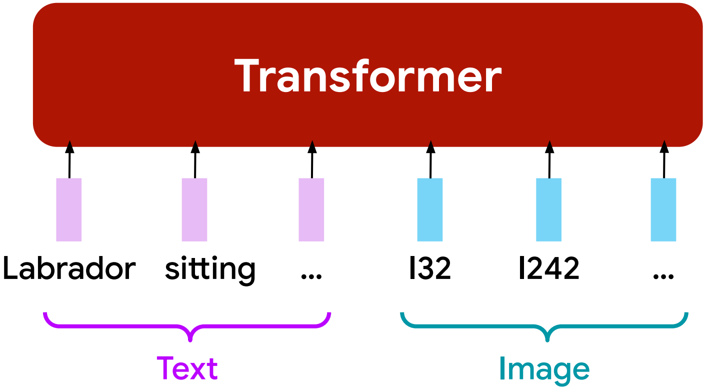
Flamingo, **GPT-4** (image → text; Transformer)



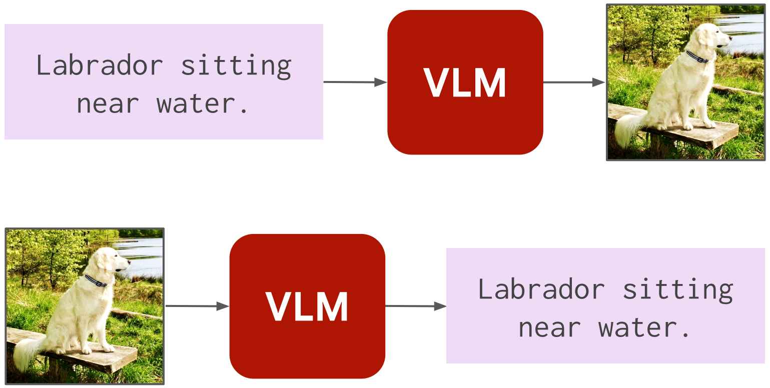
Multimodal Foundation Models (Unify Text & Image)

CM3 (text ⇄ image; Transformer)

Unified VLM



Text & image generation



Challenge

However, models may lack knowledge and **hallucinate**.

What does an **Armenian church** look like?

Text to image

✗

What is the name of this place..?

Image to text

The Dragon and Tiger Pagodas next to fireworks.

✗

Empire state building and fireworks

Challenge

Current models' knowledge is bounded by the parameters & training data. Can we allow models to **refer to external memory**?

What does an **Armenian church** look like?

Text to image

✗

What is the name of this place..?

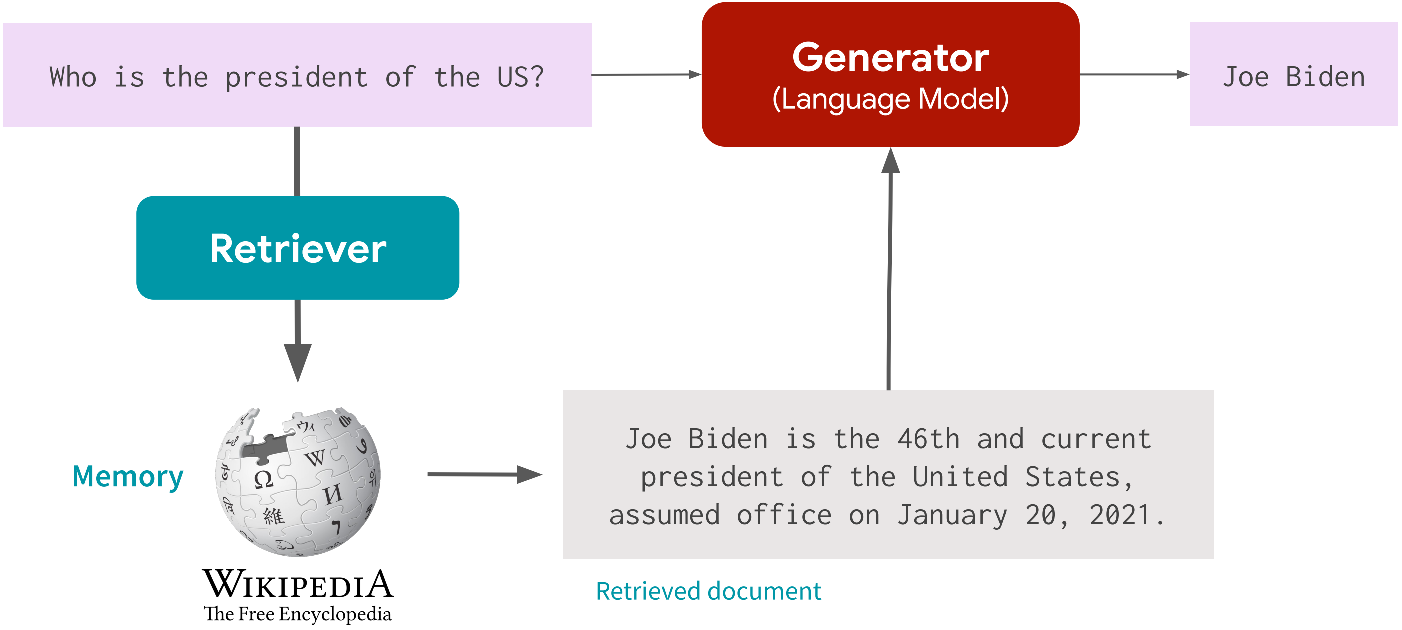
Image to text

The Dragon and Tiger Pagodas next to fireworks.

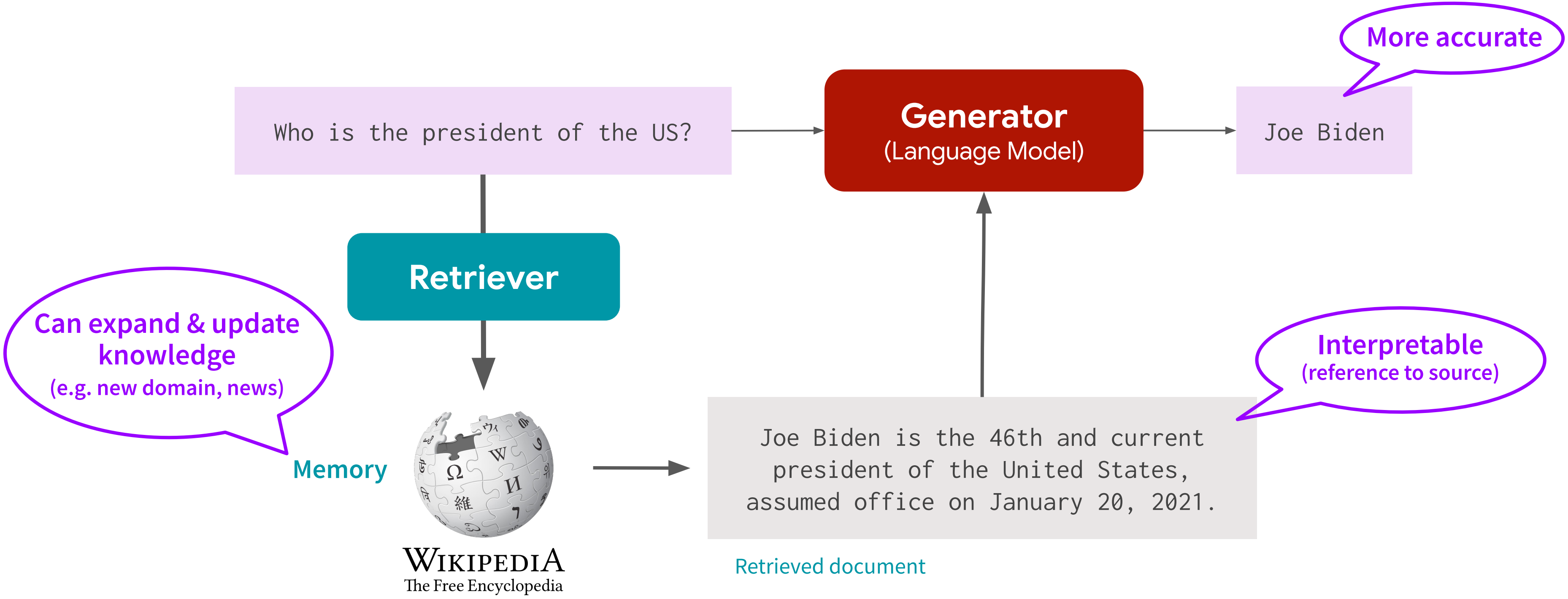
✗

Empire state building and fireworks

Inspiration: Retrieval-augmented Language Model



Inspiration: Retrieval-augmented Language Model



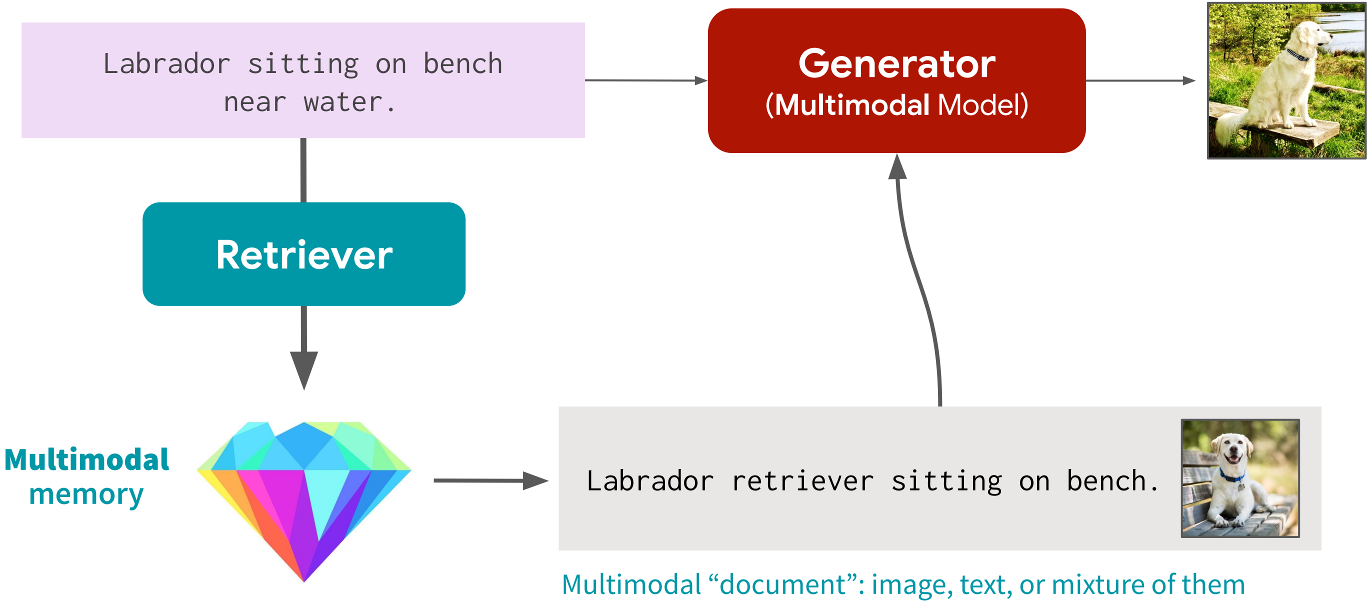
Guu+2020; Lewis+2020

Retrieval-augmented multimodal modeling

RA-CM3: Retrieval-augmented multimodal modeling.

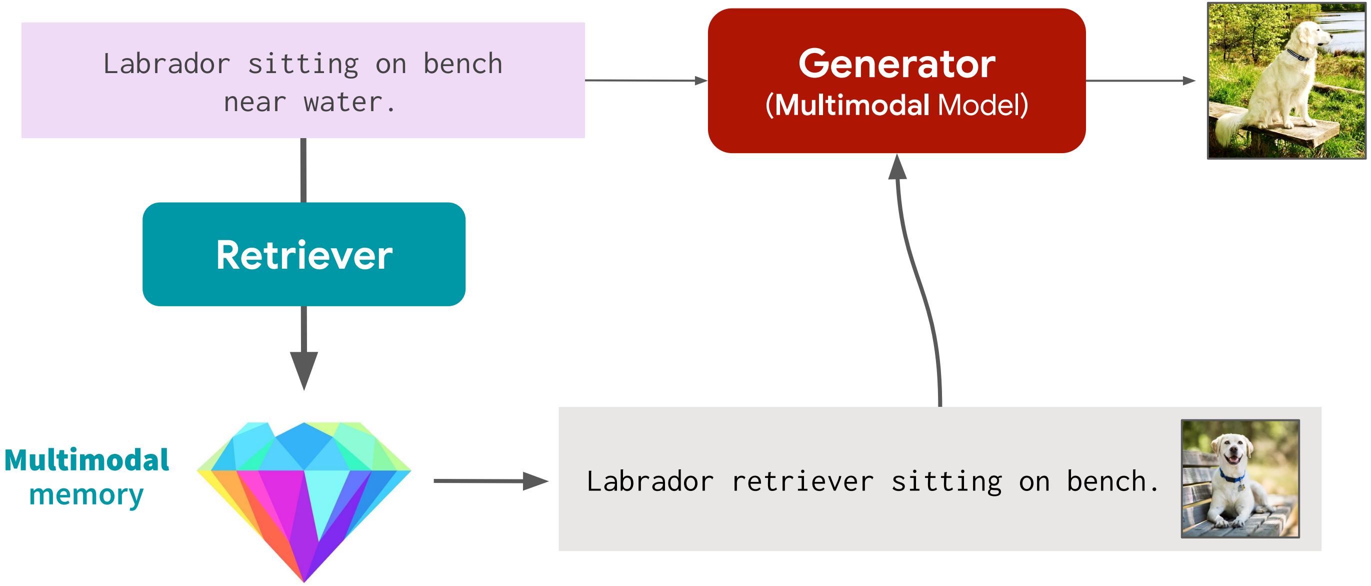
Yasunaga, Aghajanyan, Shi, James, Leskovec, Liang, Lewis, Zettlemoyer, and Yih. ICML 2023.

Our Idea: Retrieval-augmented Multimodal Model



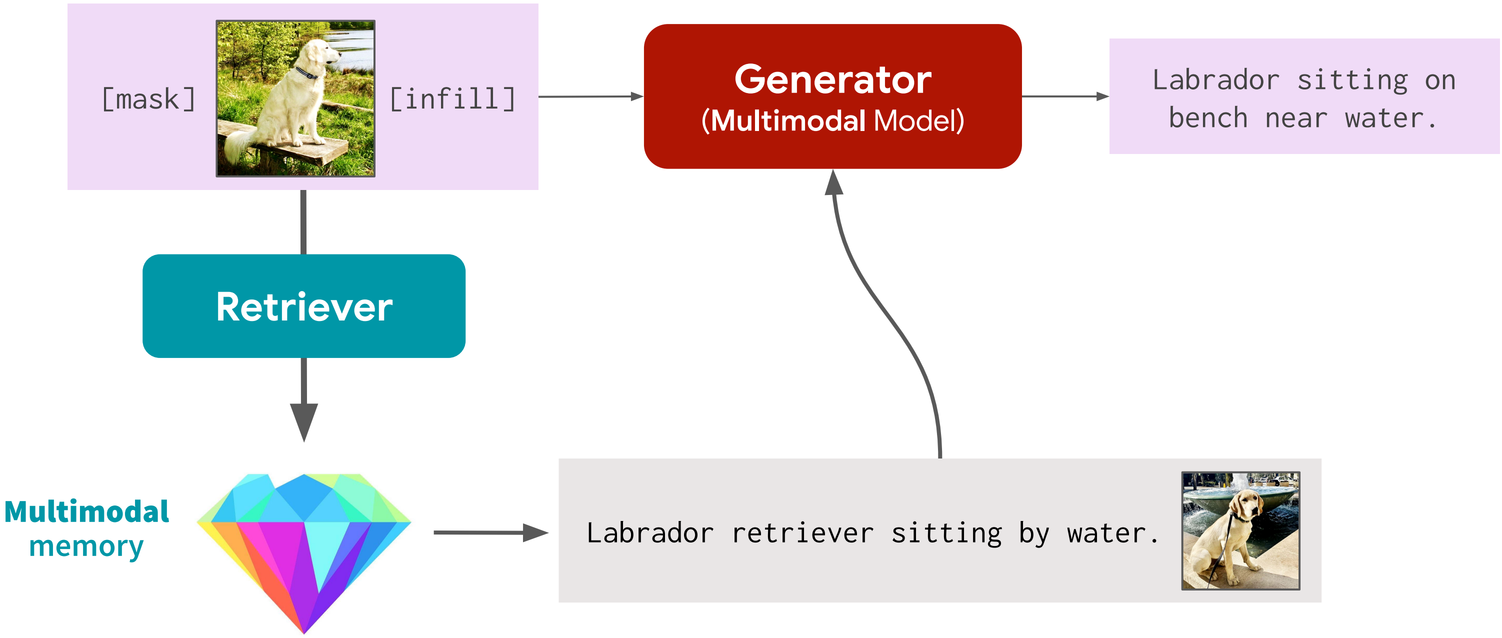
Our Idea: Retrieval-augmented Multimodal Model

Text-to-Image



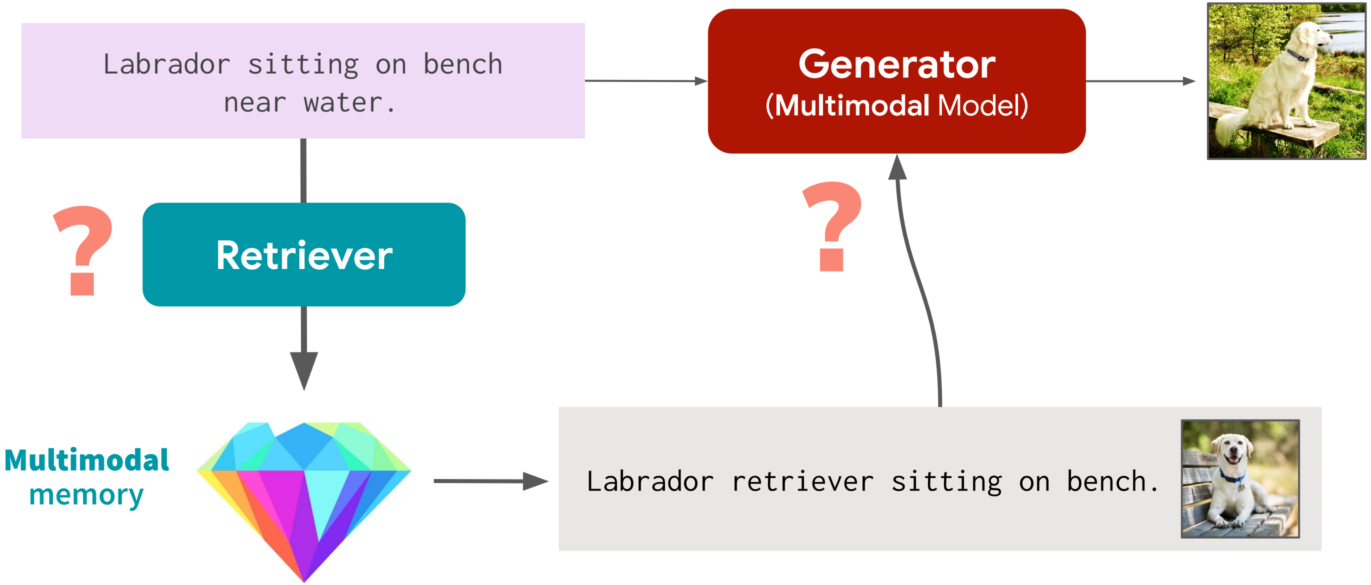
Our Idea: Retrieval-augmented Multimodal Model

Image-to-Text

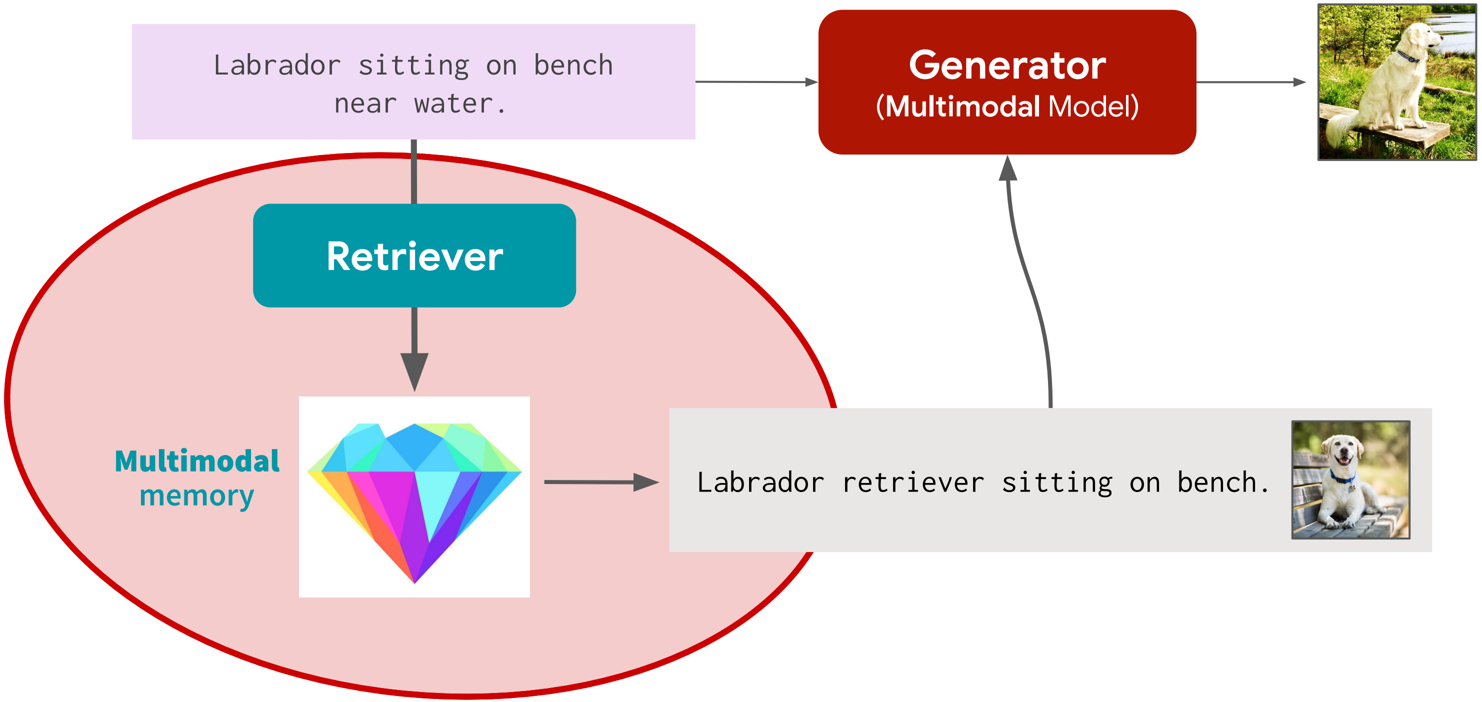


Technical innovations

- What is an effective **multimodal retrieval** method?
- How to **integrate** retrieved items into the **generator**?



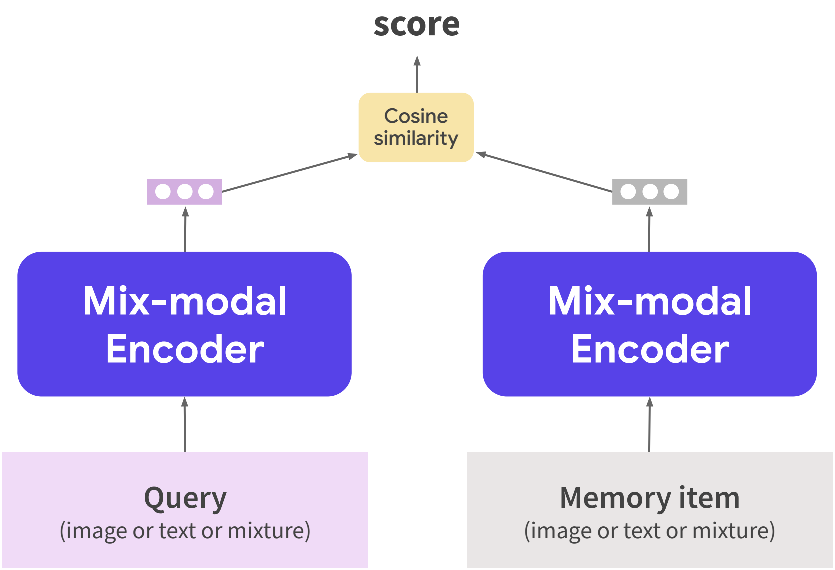
Multimodal Retrieval



Multimodal Retriever

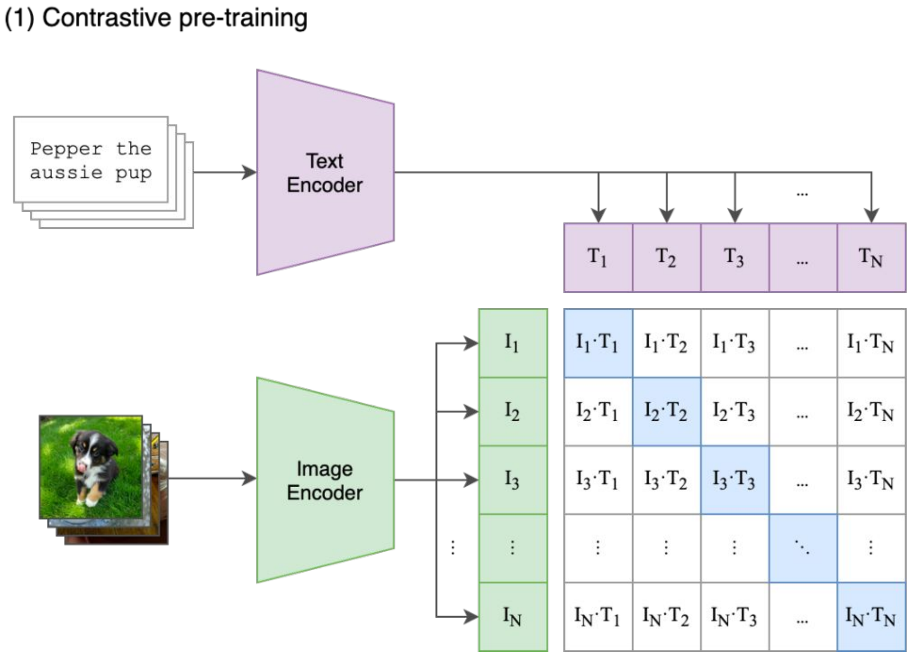
Dense Retriever with Mix-modal Encoder

`f(query, memory) ⟶ score`



Background: CLIP

CLIP produces text embeddings and image embeddings in shared vector space



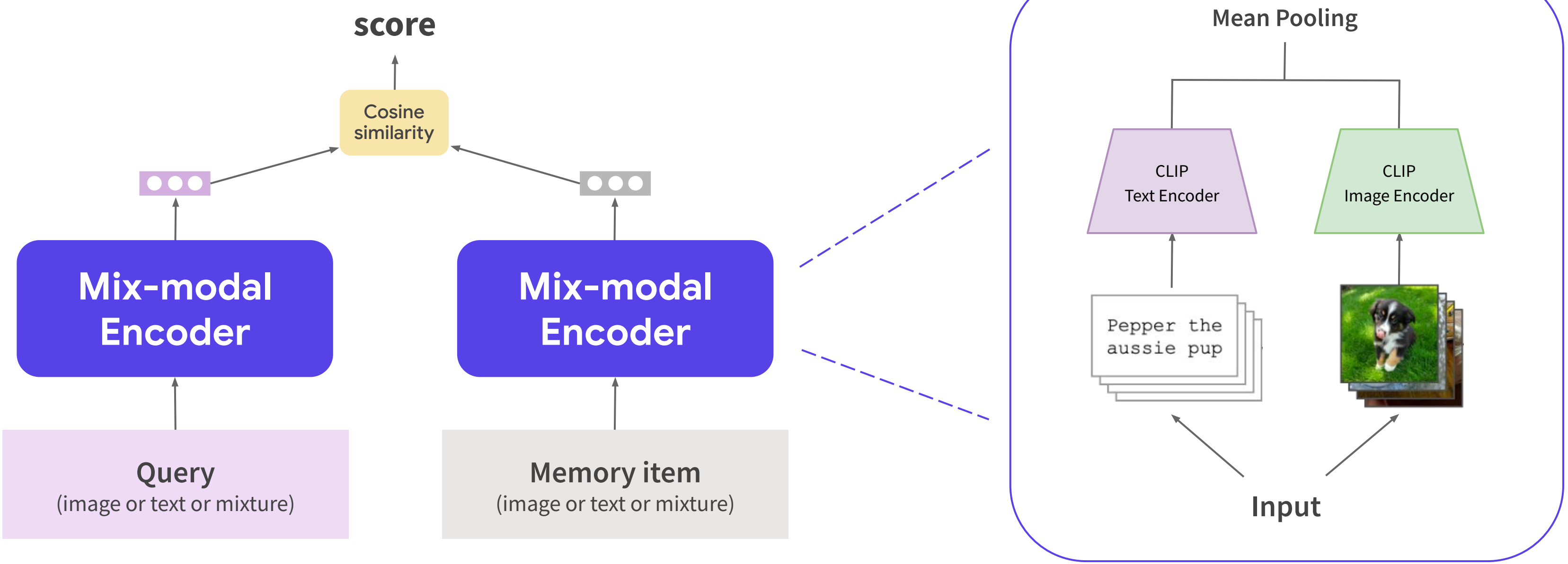
Radford+2021

Multimodal Retriever

Dense Retriever with Mix-modal Encoder

`f(query, memory) → score`

E.g. Extension of CLIP



Multimodal Retriever

Example

Memory

Query

Labrador sitting on bench near water.

0.85

Labrador retriever sitting on bench.

0.81

Labrador retriever sitting by water.

0.35

⋮

Strategy for Retrieval

Relevance

The retrieved items should be relevant to query

✅ **Cosine similarity score + Maximum Inner Product Search**

Diversity (for training)

If simply take items of top scores, may include duplicate images/text

This can cause the generator to overfit or learn repetitive generation

💡 **Avoid redundant items**

- Skip candidate item if it is too similar to query or items already retrieved

💡 **Query dropout**

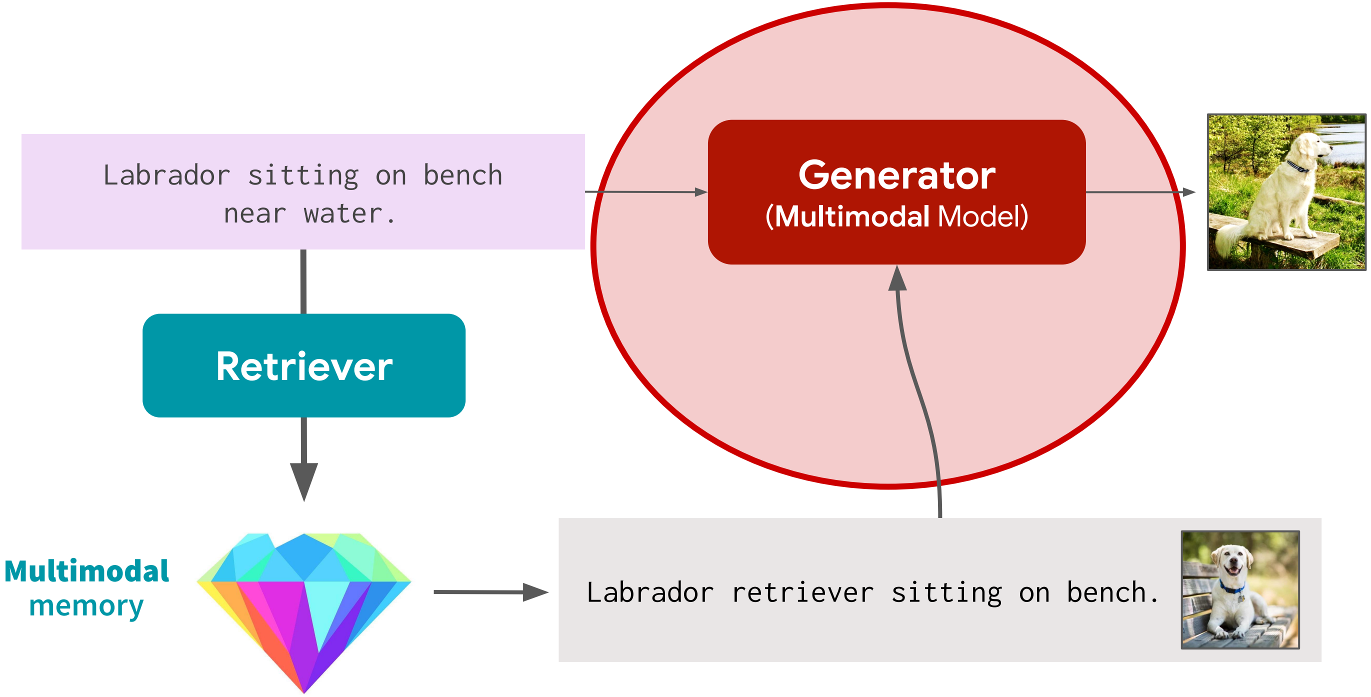
- Drop some tokens of query used in retrieval (e.g. 20% of tokens)
- This further increases diversity and serves as regularization

Diversity is crucial in multimodal setting

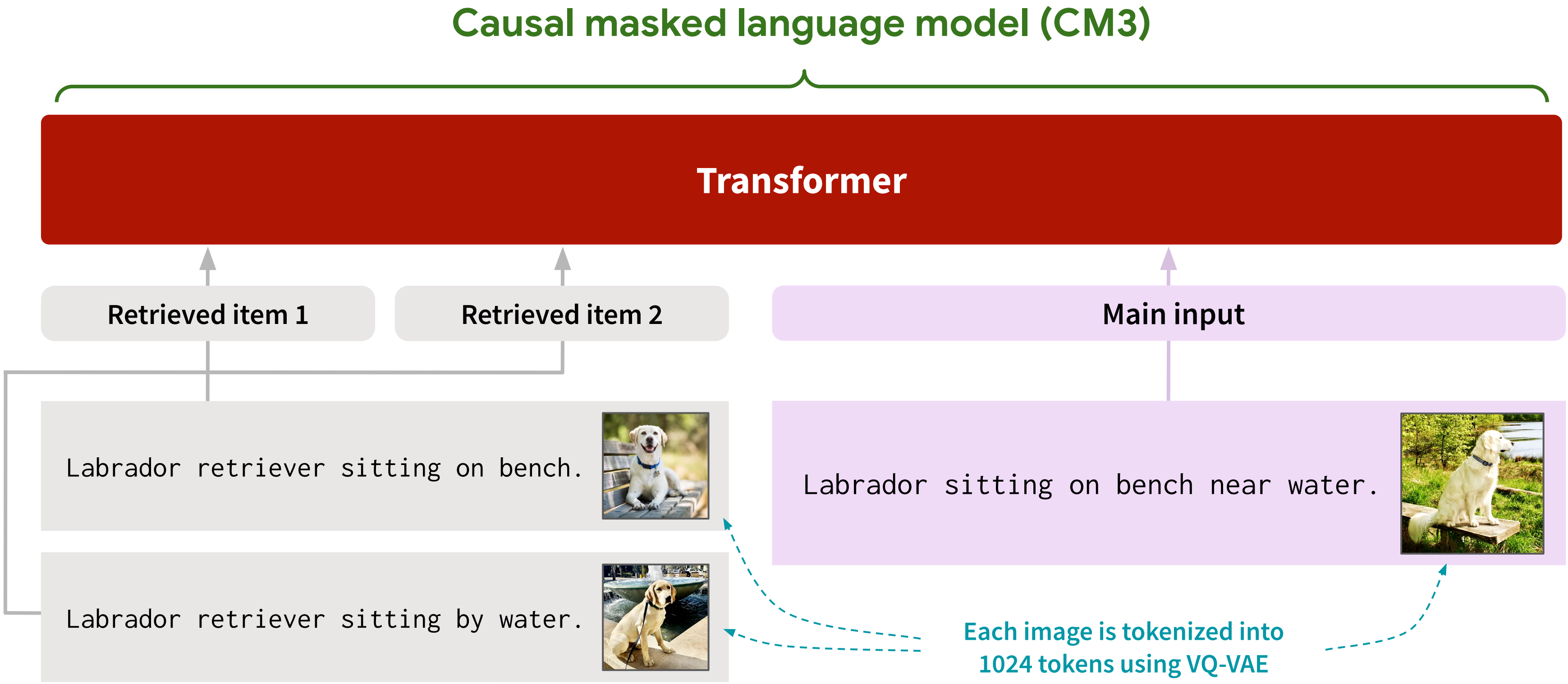
- Multimodal dataset often contains duplicate images across docs
- Each image takes many tokens (1024), so can significantly hurt model training

Can improve FID score by 5 points

Multimodal Generator



Generator: Retrieval-Augmented CM3



Train the Generator Efficiently

Loss = **(LM loss for main input)** + $\alpha \cdot$ **(LM loss for retrieved items)**

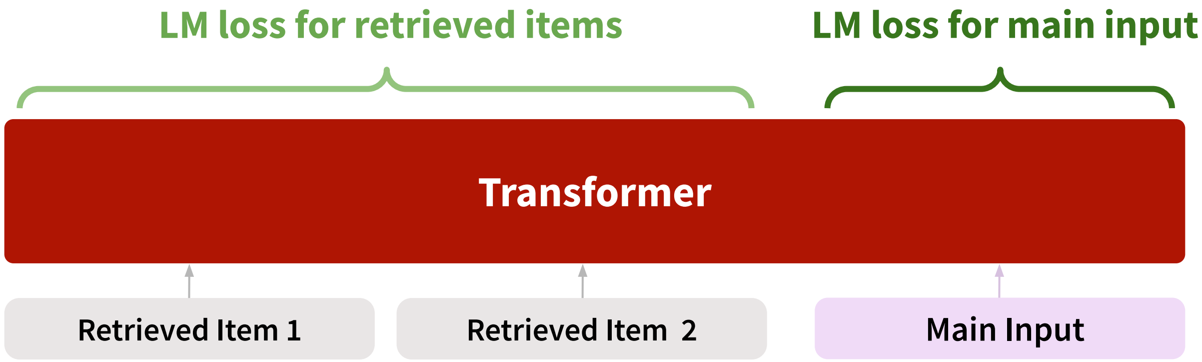
- Existing retrieval augmented LMs: $\alpha = 0$
- **Our method: $\alpha > 0$ ($\alpha = 0.1$ works the best)**

$\alpha > 0$ has effect like increasing batch size without extra forward compute, increasing training efficiency.

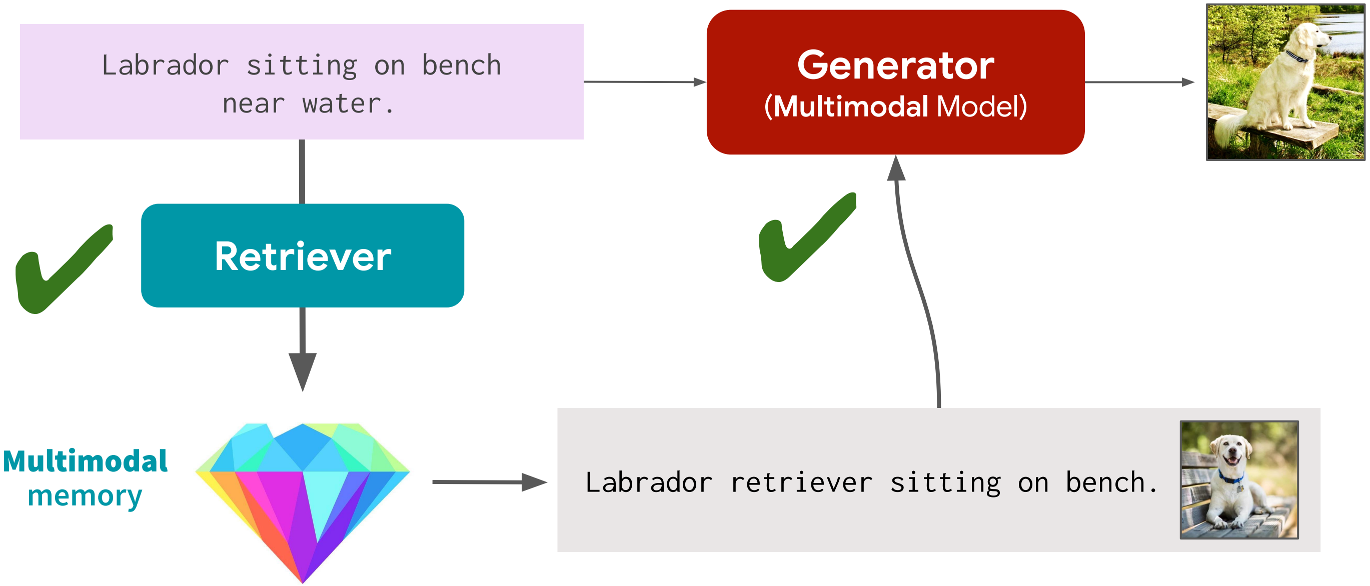
$\alpha > 0$ is crucial in multimodal setting

- Each image takes many tokens (1024)
- If $\alpha = 0$, we are throwing away a lot of compute

Can reduce training time by 50%



Retrieval Augmented Multimodal Model



Comparison with related models

Model	Image Generation	Text Generation	Retrieval
DALL-E, StableDiffusion, Imagen, etc.	✅	-	-
kNN-diffusion, Re-Imagen, etc.	✅	-	✅
Flamingo, GPT-4, etc.	-	✅	-
MuRAG, Re-ViLM, REVEAL, SmallCap, etc.	-	✅	✅
CM3	✅	✅	-
RA-CM3 (Ours)	✅	✅	✅

Experiments

Train data

- **LAION** (cleaned 150M image-text pairs)
  External memory: LAION

Evaluation

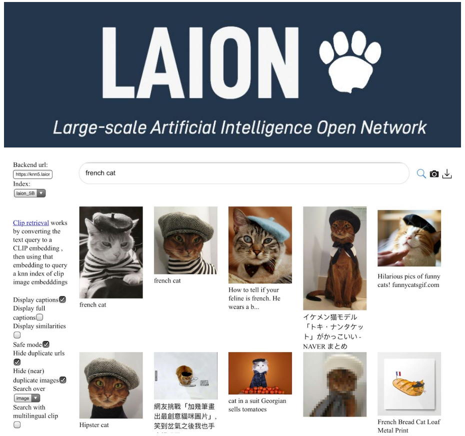
- **MSCOCO** caption2image, image2caption.
  External memory: MSCOCO train set

Model

- Transformer with seq_length 4096 (up to 2 retrieved documents)
- 2.7B parameters trained for 5 days on 256 GPUs
- **“Retrieval Augmented CM3 (RA-CM3)”**

Baseline

- Vanilla CM3 with no retrieval, same size, trained using the same amount of compute



Performance (Text-to-Image)

Retrieval improves caption-to-image generation quality (e.g. RA-CM3 vs CM3)

Model	Model type	#Train images	MSCOCO FID score (↓)
DALL-E (12B)	Autoregressive	250M	28
Parti (20B)	Autoregressive	6B	7.2
Stable Diffusion	Diffusion	1B	~12
Vanilla CM3	Autoregressive	150M	29
RA-CM3	Autoregressive	150M	**16**

13 points improvement

(`An Armenian church.` → Text to image )

Performance (Image-to-Text)

Retrieval improves image-to-caption generation quality (e.g. RA-CM3 vs CM3)

Model	#Train images	MSCOCO CIDEr score (↑)
Parti (20B)	6B	0.84
Flamingo (3B) 4-shot	2.5B	0.85
Vanilla CM3	150M	0.72
RA-CM3	150M	**0.89**

17 points improvement

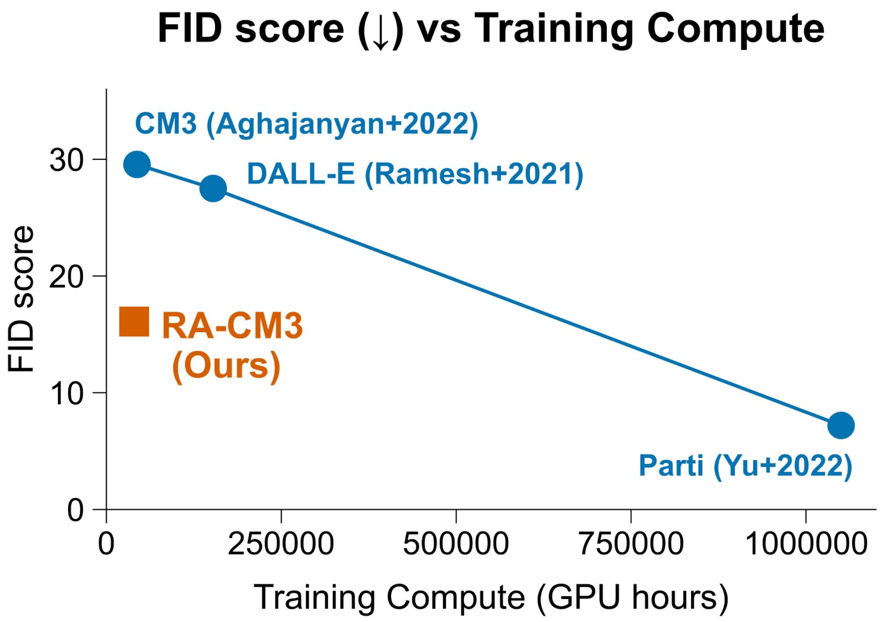
Image to text

The Dragon and Tiger Pagodas next to fireworks.

Performance (Efficiency)

Retrieval improves training efficiency

- RA-CM3 outperforms DALL-E while using only 30% of training compute



Accurate Image Generation

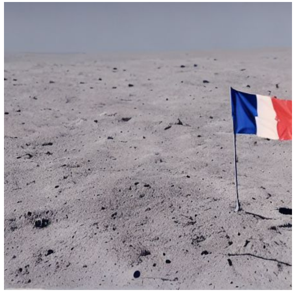
RA-CM3 Retrieved items

RA-CM3 outputs

Baseline outputs

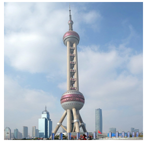
(Vanilla CM3) (Stable Diffusion)

French flag



Input: "**French flag** waving on the moon's surface."

Oriental Pearl tower



Input: "The **Oriental Pearl tower** in oil painting."

Accurate Image Generation

RA-CM3 Retrieved items

RA-CM3 outputs

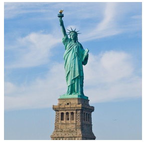
Baseline outputs

(Vanilla CM3) (Stable Diffusion)

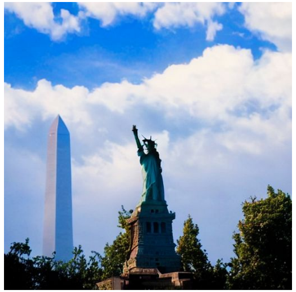
Armenian church

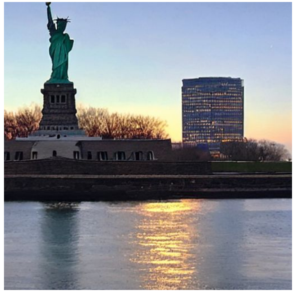
Input: "An **Armenian church** during a sunny day."

Statue of Liberty



Washington monument





Input: "Photo of the **Statue of Liberty** standing next to the **Washington monument**."

Accurate Image Generation

RA-CM3 Retrieved items

RA-CM3 outputs

Baseline outputs

(Vanilla CM3) (Stable Diffusion)

Ming Dynasty vase

Input: "A **Ming Dynasty vase** with orange flowers painted."

Mount Rushmore

Japanese cherry

Input: "The **Mount Rushmore** with **Japanese cherry** trees in the front."

Accurate Image Generation

RA-CM3 Retrieved items

RA-CM3 outputs

Baseline outputs

(Vanilla CM3) (Stable Diffusion)

Callanish standing stones

Input: “Photo of the **Callanish standing stones**, fireworks in the sky.”

Dragon and Tiger Pagodas

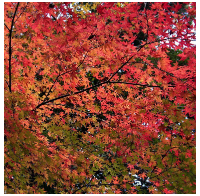
Input: “Photo of **the Dragon and Tiger Pagodas**, the sun is setting behind.”

Multimodal In-Context Learning

RA-CM3 In-context | **RA-CM3 output** | **Baseline outputs**

(Vanilla CM3) (Stable Diffusion)

(Demonstrate the style to generate)


"Photo of a house taken on an autumn day."

(Demonstrate the style to generate)





"Painting of red roses."

Intuition:
After retrieval augmented training, our generator model has learned how to use in-context examples and acquired this in-context learning capability

Image Editing

Provide an image to control the type of editing

Source image

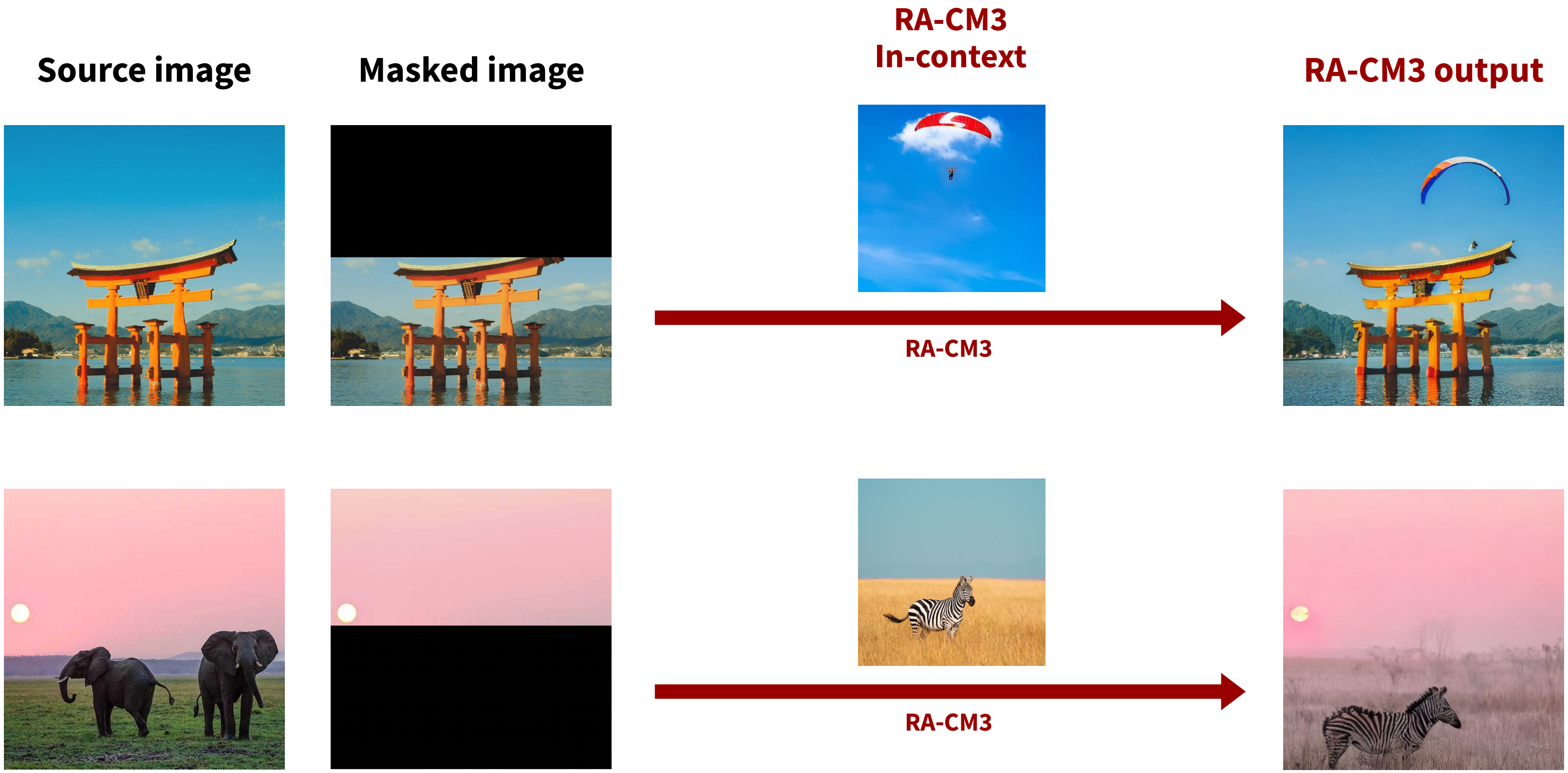
Masked image

RA-CM3
In-context

RA-CM3

RA-CM3 output

Image Editing



One-shot Image-to-Text

Task

 animal X  animal Y  animal __

RA-CM3

$P(\mathrm{X}),\ P(\mathrm{Y})$

Binary classification

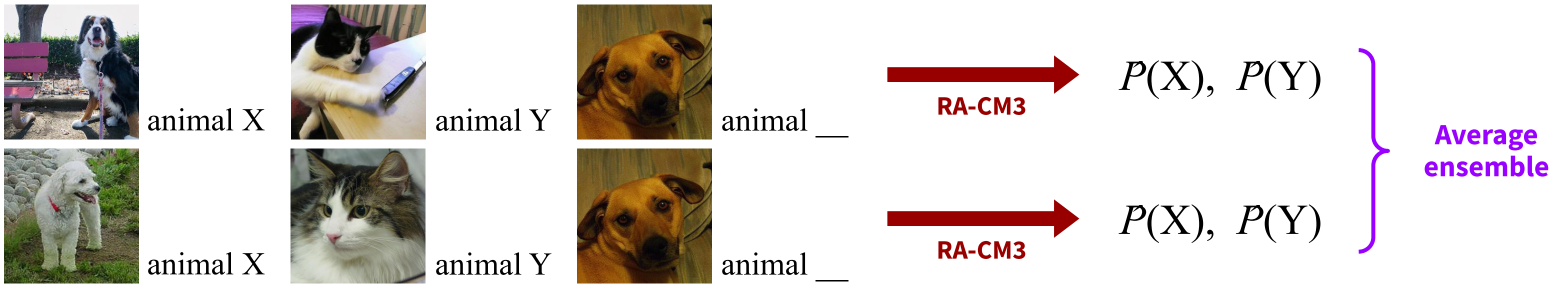
Result

Model	Accuracy
Baseline CM3	0.53
RA-CM3	**0.78**

Motivation: test the true in-context learning capability of our generator

Few-shot Image-to-Text

Ensemble (e.g. 2)



Result

Model	Number of Ensembles			
	1	2	4	8
Baseline CM3	0.53	0.50	0.56	0.56
RA-CM3	**0.78**	**0.79**	**0.86**	**0.9**

Takeaway:

- Generator exhibits good in-context learning performance
- Ensemble is an effective method to increase in-context examples

Summary

RA-CM3: The first retrieval-augmented multimodal model that can retrieve and generate both text and images

Result & Impact: Retrieval enables

- Accurate image/text generation ⇒ **reduce hallucination**
- Efficient training ⇒ **reduce cost** of training large foundation models
- Multimodal in-context learning (e.g., can prompt using both images and text)

Thak you!

https://cs.stanford.edu/~myasu/

@michiyasunaga