Grounding Language Models to Images for Multimodal Inputs and Outputs

Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried

ICML 2023

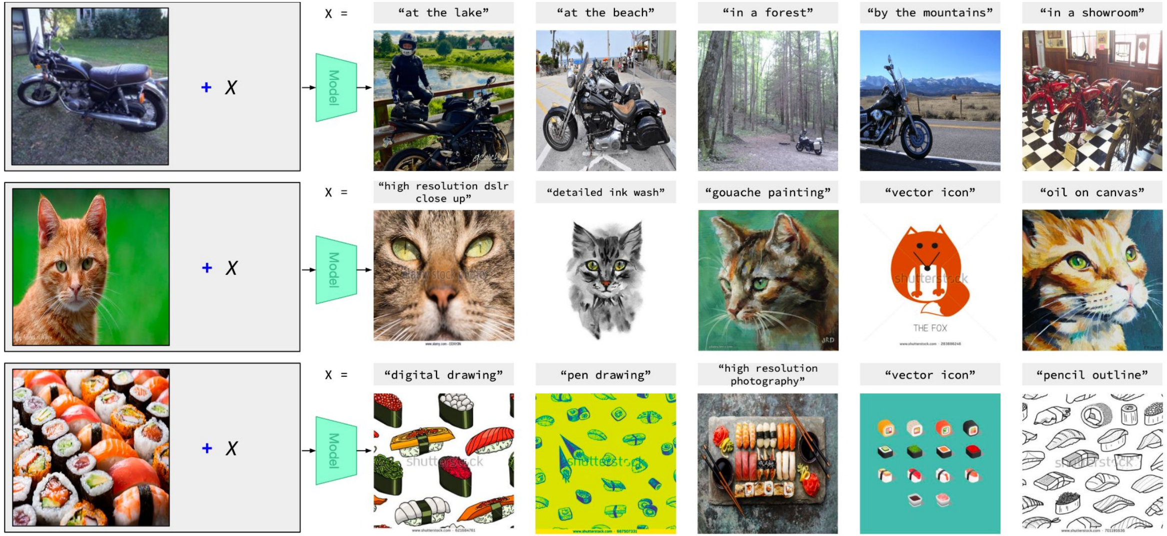
Carnegie Mellon University

Can we ground text-only LLMs to consume and produce visual data?

🧀 FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

jykoh.com/fromage



Concept composition. Our model is capable of retrieving relevant images conditioned on multi-modal context inputs.

Grounding Language Models to Images for Multimodal Generation (jykoh.com/fromage)

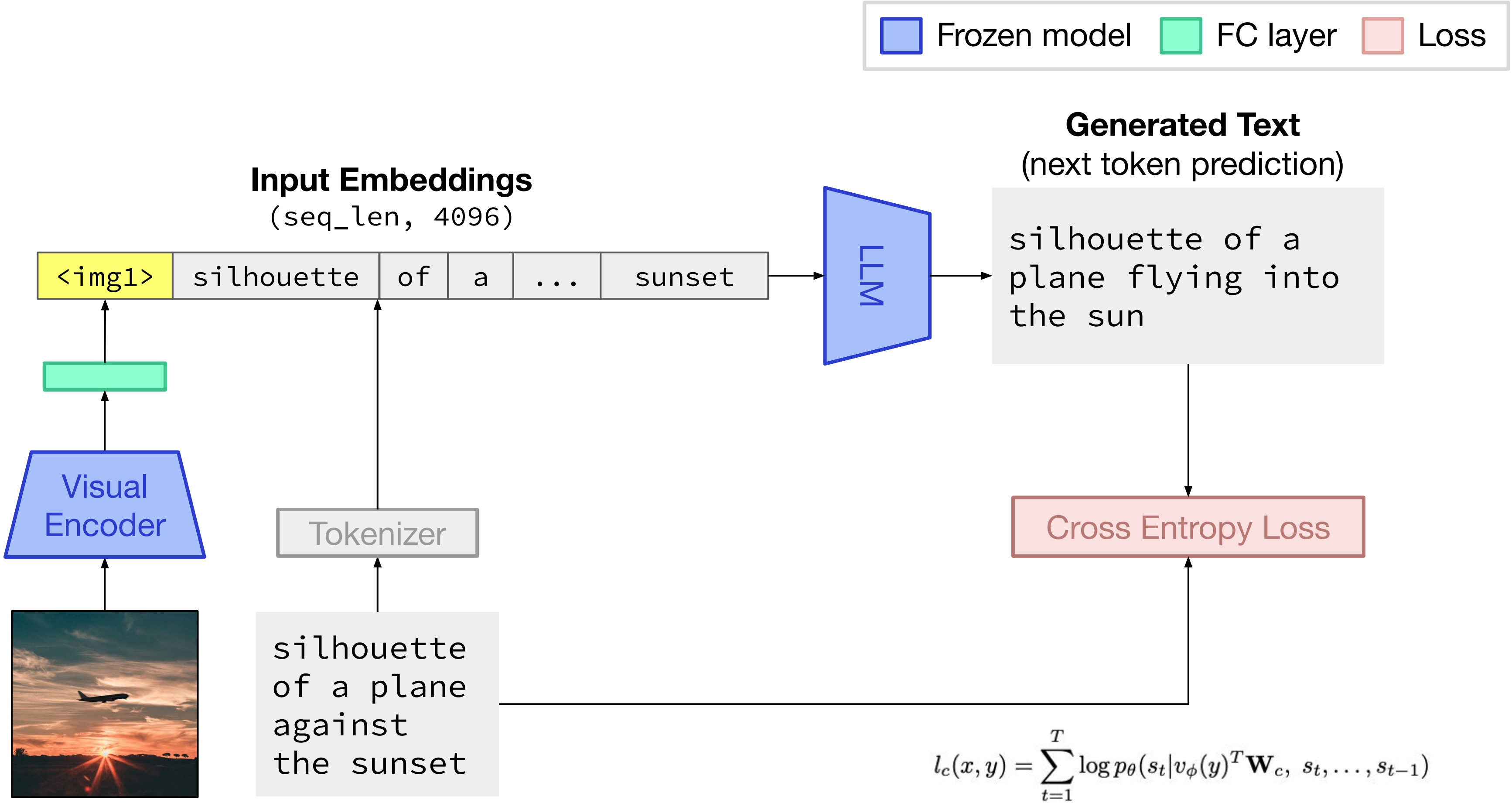
Multi-modal dialogue. Green bubbles represent model generated outputs, grey bubbles represent user input.

🧀 FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

- **Leverage the learnt abilities of pre-trained text-only LLMs**
  - In-context learning
  - Sensitivity to input prompts
  - Generate long and coherent dialogue
- **Model agnostic**
  - We use a 6.7B LLM (past the scale necessary for generalization to larger models)
  - Can (in principle) be applied to any larger model, and any stronger LLM released in the future
- **Simple and resource efficient**
  - We train just 3 linear layers to adapt a text-only LLM for image captioning + image retrieval
  - FROMAGe is trained on a single A6000 GPU in 24 hours

Image Captioning



🟦 Frozen model 🟩 FC layer 🟥 Loss

Input Embeddings

`(seq_len, 4096)`

Generated Text

(next token prediction)

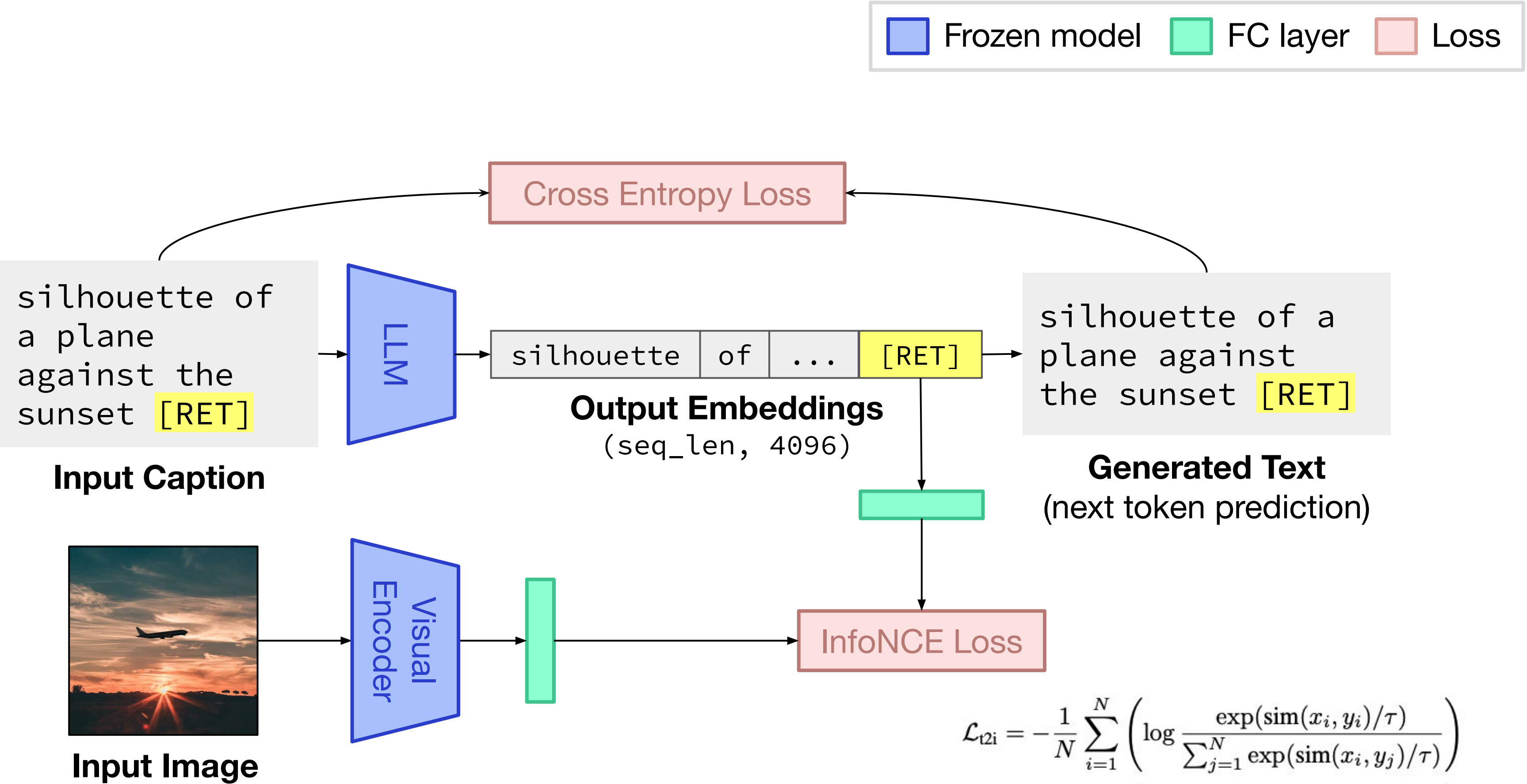
`silhouette of a plane flying into the sun`

`silhouette of a plane against the sunset`

Image #1

Caption #1

$$l_c(x,y) = \sum_{t=1}^{T} \log p_\theta(s_t | v_\phi(y)^T \mathbf{W}_c,\ s_t, \ldots, s_{t-1})$$



$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x_i, y_j)/\tau)}\right)$$

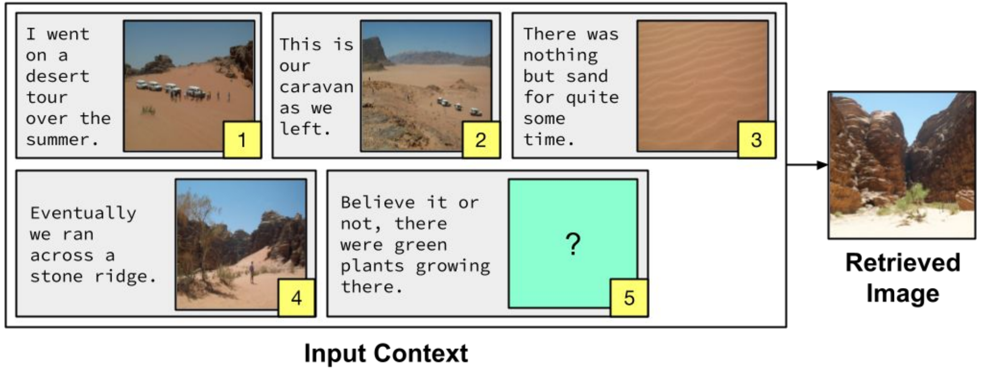
$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{\exp(\text{sim}(y_i, x_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(y_i, x_j)/\tau)}\right)$$

Image-Text Retrieval

Quantitative Evaluations

1) Contextual image retrieval

- Given a Visual Story, retrieve the correct image
- FROMAGe is more sensitive to context
- CLIP gets worse with more context



I went on a desert tour over the summer.

This is our caravan as we left.

There was nothing but sand for quite some time.

Eventually we ran across a stone ridge.

Believe it or not, there were green plants growing there.

Model	Inputs	R@1	R@5	R@10
CLIP ViT-L/14	1 caption	**11.9**	**25.5**	**32.2**
FROMAGe		9.0	21.1	28.7
CLIP ViT-L/14	5 captions	5.9	19.5	28.0
FROMAGe		**10.4**	**23.8**	**31.7**
BLIP[†]	5 captions	6.2	16.8	23.4
CLIP ViT-L/14[†]	5 captions	8.8	22.3	29.8
FROMAGe [†]	5 captions	11.6	24.7	32.8
CLIP ViT-L/14	5 captions, 4 images	2.4	21.3	34.0
FROMAGe [†]	5 captions, 4 images	**15.6**	**36.5**	**45.8**

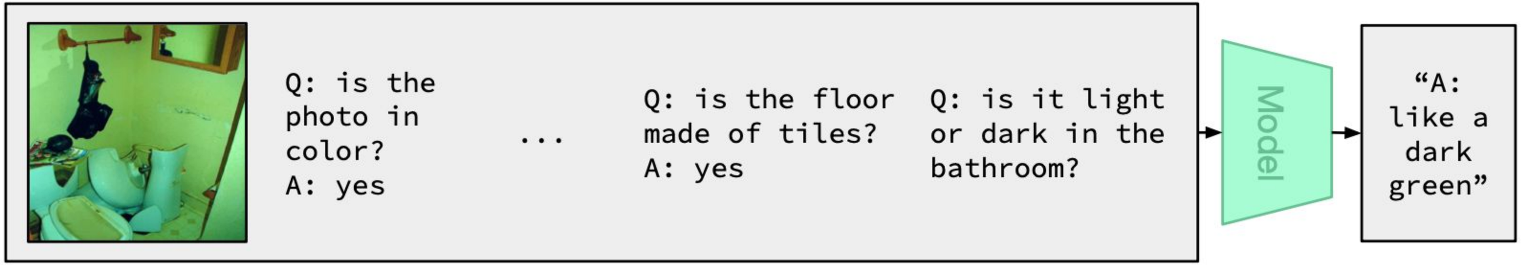
Table 1. Recall@k on zero-shot contextual image retrieval of the last image in Visual Storytelling (Huang et al., 2016). Numbers in **bold** indicate best scores for a particular set of inputs. [†] indicates retrieval over images not previously seen in the story sequence.

Visual Storytelling (Huang et al., 2016)

Quantitative Evaluations

2) Visual Dialogue

- **IT2T:** Answer a question about the image given past dialogue discussing it



Q: is the photo in color?
A: yes

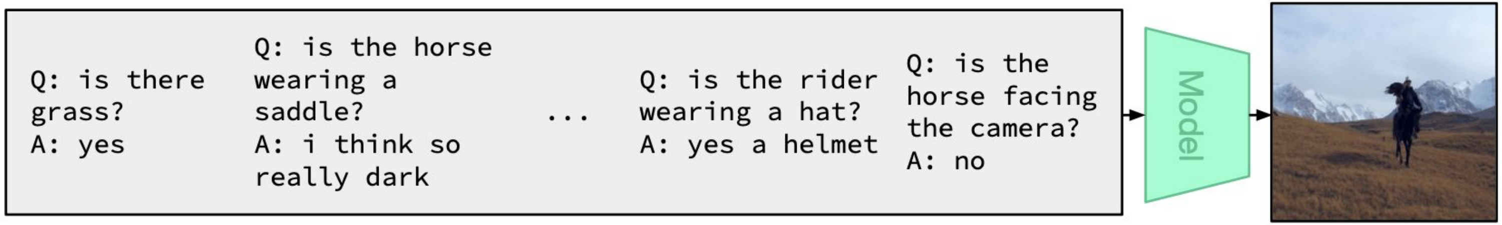
...

Q: is the floor made of tiles?
A: yes

Q: is it light or dark in the bathroom?

“A: like a dark green”

- **T2I:** Retrieve the correct image given a series of dialogue about it



Q: is there grass?
A: yes

Q: is the horse wearing a saddle?
A: i think so really dark

...

Q: is the rider wearing a hat?
A: yes a helmet

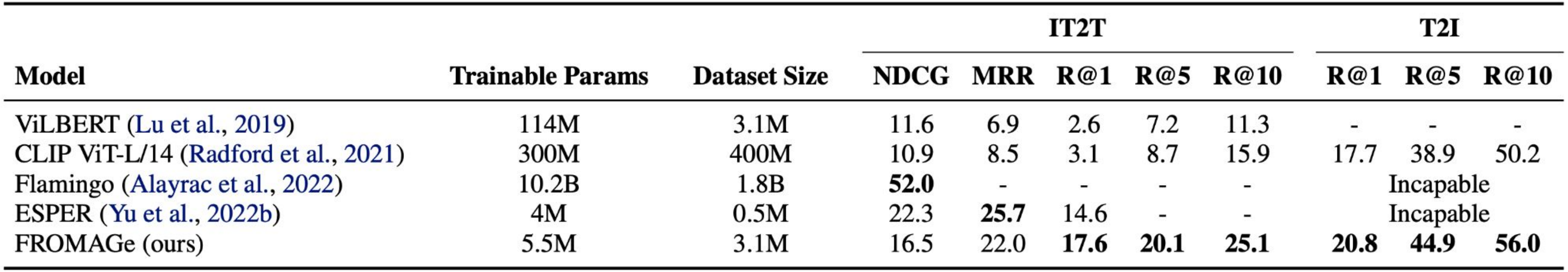
Q: is the horse facing the camera?
A: no

Visual Dialog (Das et al., 2017)

Quantitative Evaluations

2) Visual Dialogue

- **IT2T:** Answer a question about the image given past dialogue discussing it
- **T2I:** Retrieve the correct image given a series of dialogue about it



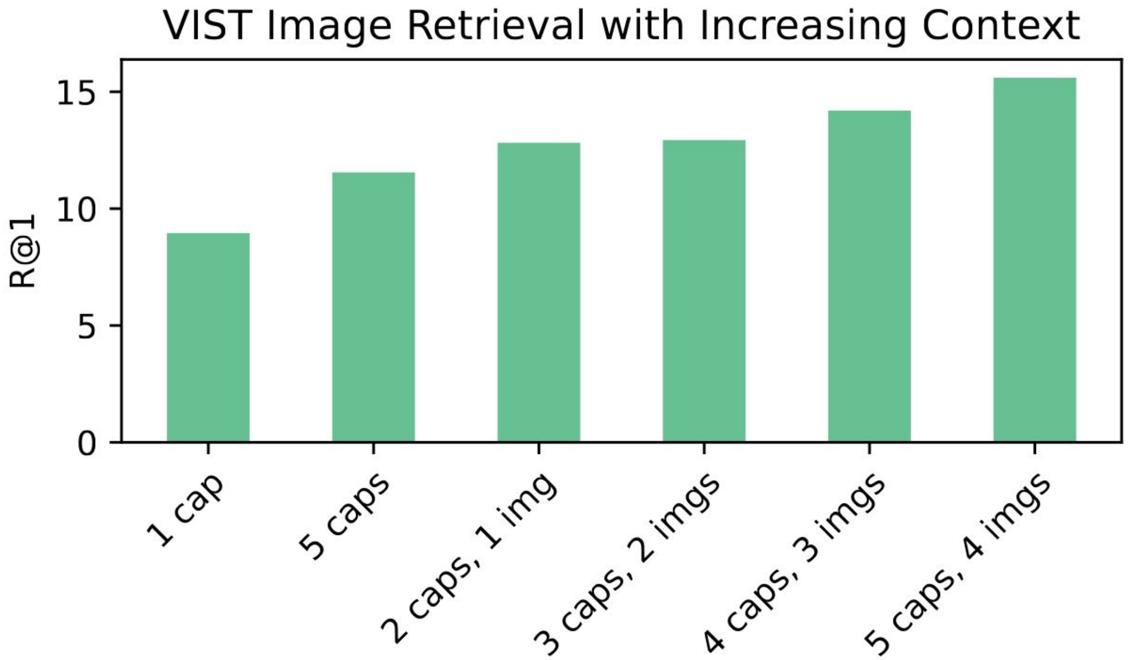
			IT2T					T2I		
Model	**Trainable Params**	**Dataset Size**	**NDCG**	**MRR**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
ViLBERT (Lu et al., 2019)	114M	3.1M	11.6	6.9	2.6	7.2	11.3	-	-	-
CLIP ViT-L/14 (Radford et al., 2021)	300M	400M	10.9	8.5	3.1	8.7	15.9	17.7	38.9	50.2
Flamingo (Alayrac et al., 2022)	10.2B	1.8B	**52.0**	-	-	-	-		Incapable	
ESPER (Yu et al., 2022b)	4M	0.5M	22.3	**25.7**	14.6	-	-		Incapable	
FROMAGe (ours)	5.5M	3.1M	16.5	22.0	**17.6**	**20.1**	**25.1**	**20.8**	**44.9**	**56.0**

Table 2. Zero-shot results on Visual Dialog (Das et al., 2017), for image-and-text-to-text (IT2T) and text-to-image (T2I) retrieval. Unlike previous methods, FROMAGe is capable of generating free-form text interleaved with image outputs through text-to-image retrieval.

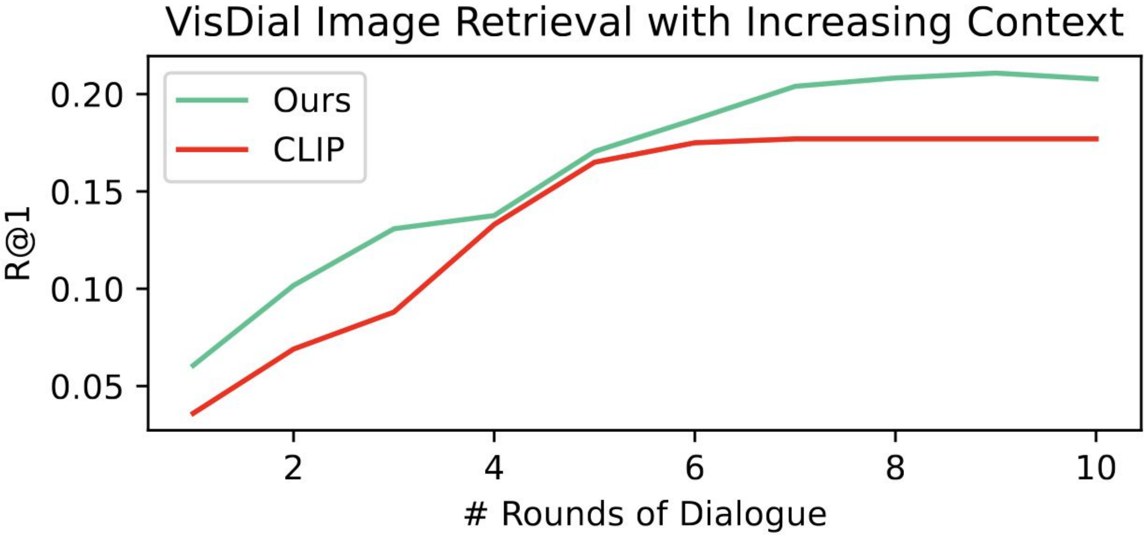
Visual Dialog (Das et al., 2017)

The Effect of Context

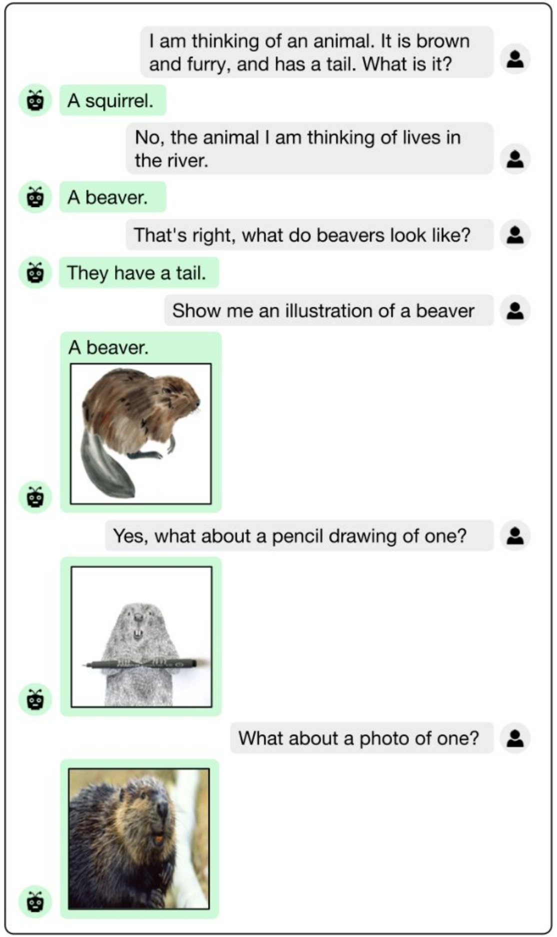
Multi-modal context is **worth more** than uni-modal context.



The Effect of Context



Beavers



I am thinking of an animal. It is brown and furry, and has a tail. What is it?

A squirrel.

No, the animal I am thinking of lives in the river.

A beaver.

That's right, what do beavers look like?

They have a tail.

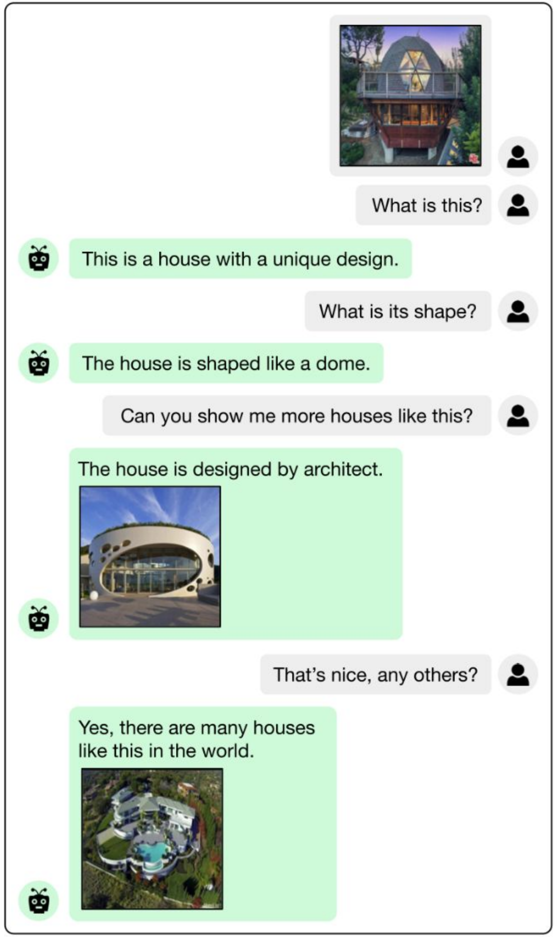
Show me an illustration of a beaver

A beaver.

Yes, what about a pencil drawing of one?

What about a photo of one?

Unique Houses



What is this?

This is a house with a unique design.

What is its shape?

The house is shaped like a dome.

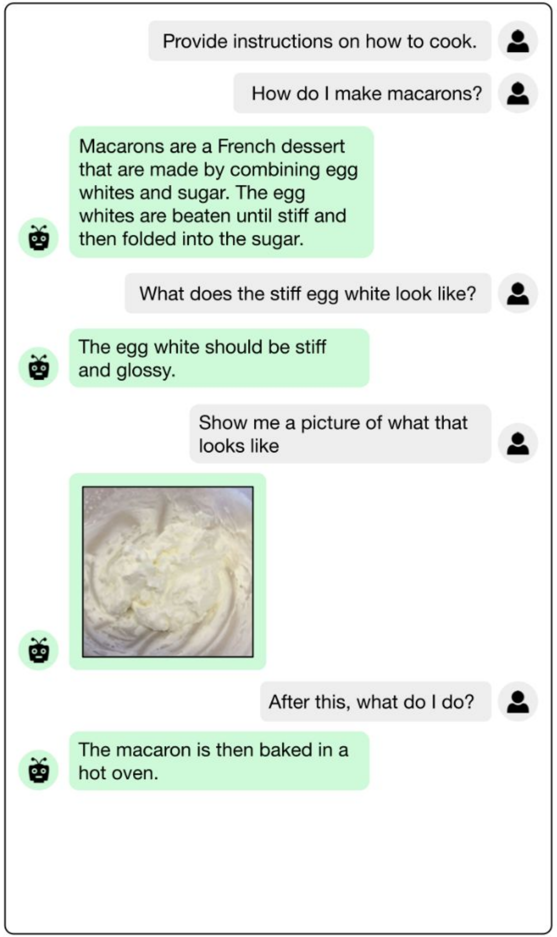
Can you show me more houses like this?

The house is designed by architect.

That's nice, any others?

Yes, there are many houses like this in the world.

How to Make Macarons



Provide instructions on how to cook.

How do I make macarons?

Macarons are a French dessert that are made by combining egg whites and sugar. The egg whites are beaten until stiff and then folded into the sugar.

What does the stiff egg white look like?

The egg white should be stiff and glossy.

Show me a picture of what that looks like

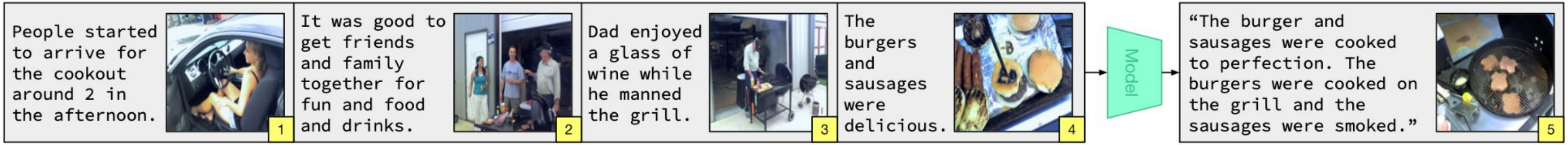
After this, what do I do?

The macaron is then baked in a hot oven.

Multi-Modal Dialogue

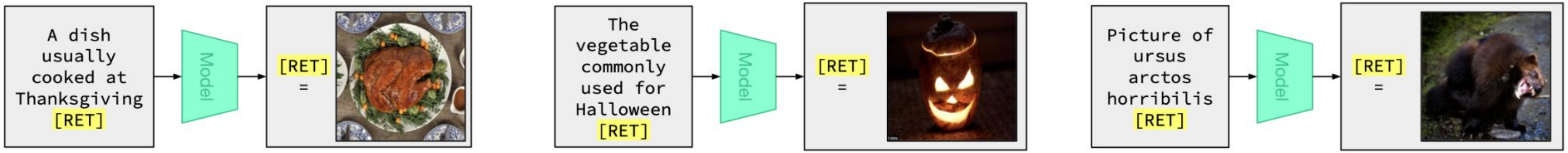
Green bubbles indicate model generated outputs, grey bubbles indicate user provided prompts.

In-Context Learning and Other Abilities


People started to arrive for the cookout around 2 in the afternoon.
1
It was good to get friends and family together for fun and food and drinks.
2
Dad enjoyed a glass of wine while he manned the grill.
3
The burgers and sausages were delicious.
4
Model
"The burger and sausages were cooked to perfection. The burgers were cooked on the grill and the sausages were smoked."
5

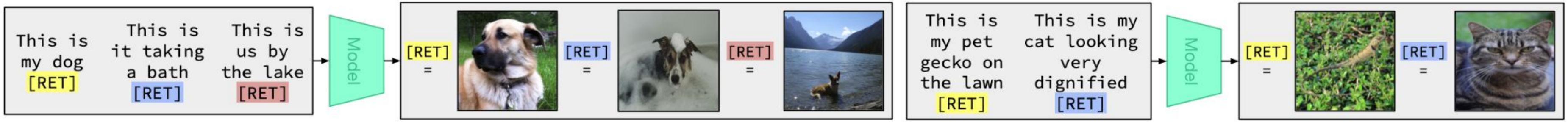
Visual Storytelling

FROMAGe can learn from in-context examples to generate story-like image and text outputs.


A dish usually cooked at Thanksgiving [RET]
Model
[RET] =
The vegetable commonly used for Halloween [RET]
Model
[RET] =
Picture of ursus arctos horribilis [RET]
Model
[RET] =

World Knowledge

Our method can draw upon knowledge learnt outside of CC3M (through large scale text pretraining of the frozen LLM) to return valid image outputs.


This is my dog [RET]
This is it taking a bath [RET]
This is us by the lake [RET]
Model
[RET] =
[RET] =
[RET] =
This is my pet gecko on the lawn [RET]
This is my cat looking very dignified [RET]
Model
[RET] =
[RET] =

Interleaved Text-to-Image Composition

Our model can transform a sequence of text inputs into text-and-image outputs. It can do coreferencing to select the appropriate images.

Thanks!

jykoh@cmu.edu
jykoh.com/fromage