

Grounding Language Models to Images for Multimodal Inputs and Outputs

Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried

22 March 2023

London Machine Learning Meetup

The logo of Carnegie Mellon University, featuring the text "Carnegie Mellon University" in white serif font on a red square background.

Carnegie
Mellon
University

LLMs are impressive general models

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

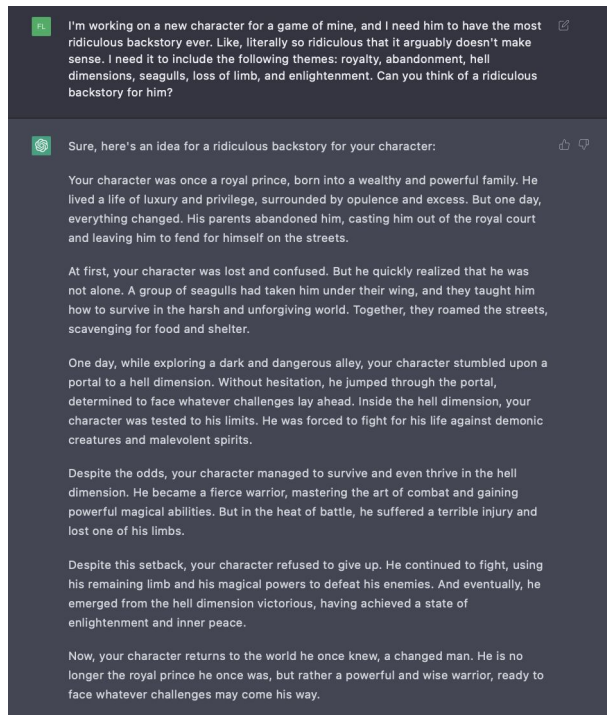
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓



I'm working on a new character for a game of mine, and I need him to have the most ridiculous backstory ever. Like, literally so ridiculous that it arguably doesn't make sense. I need it to include the following themes: royalty, abandonment, hell dimensions, seagulls, loss of limb, and enlightenment. Can you think of a ridiculous backstory for him?

Sure, here's an idea for a ridiculous backstory for your character:

Your character was once a royal prince, born into a wealthy and powerful family. He lived a life of luxury and privilege, surrounded by opulence and excess. But one day, everything changed. His parents abandoned him, casting him out of the royal court and leaving him to fend for himself on the streets.

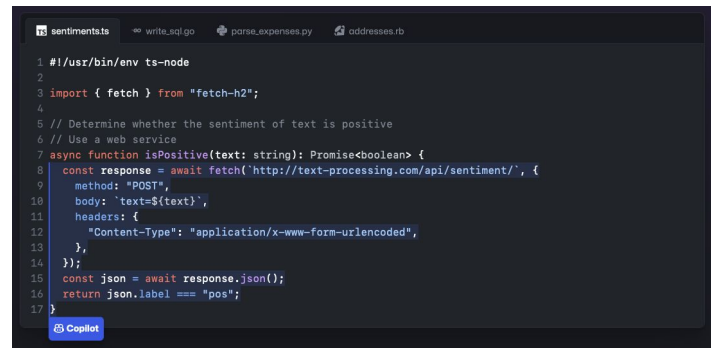
At first, your character was lost and confused. But he quickly realized that he was not alone. A group of seagulls had taken him under their wing, and they taught him how to survive in the harsh and unforgiving world. Together, they roamed the streets, scavenging for food and shelter.

One day, while exploring a dark and dangerous alley, your character stumbled upon a portal to a hell dimension. Without hesitation, he jumped through the portal, determined to face whatever challenges lay ahead. Inside the hell dimension, your character was tested to his limits. He was forced to fight for his life against demonic creatures and malevolent spirits.

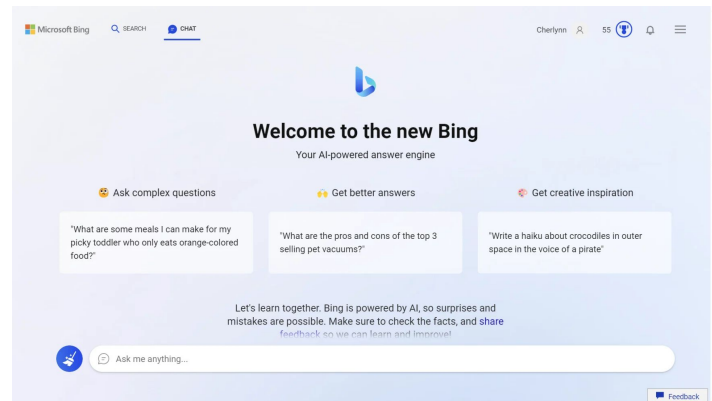
Despite the odds, your character managed to survive and even thrive in the hell dimension. He became a fierce warrior, mastering the art of combat and gaining powerful magical abilities. But in the heat of battle, he suffered a terrible injury and lost one of his limbs.

Despite this setback, your character refused to give up. He continued to fight, using his remaining limb and his magical powers to defeat his enemies. And eventually, he emerged from the hell dimension victorious, having achieved a state of enlightenment and inner peace.

Now, your character returns to the world he once knew, a changed man. He is no longer the royal prince he once was, but rather a powerful and wise warrior, ready to face whatever challenges may come his way.



```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch('http://text-processing.com/api/sentiment/', {
9     method: "POST",
10    body: 'text=${text}',
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17 }
```



Chain of Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

ChatGPT (OpenAI, 2022)

Copilot (GitHub, 2021)

Bing (Microsoft, 2023)

LLMs are impressive general models

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

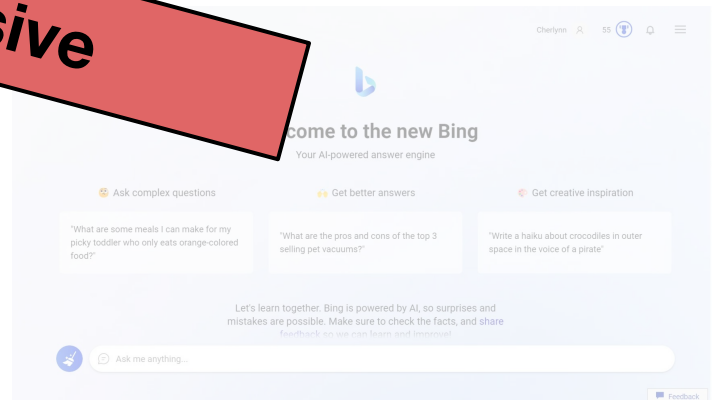
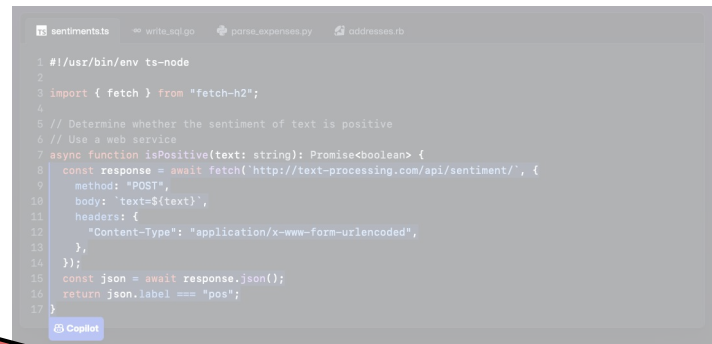
(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Resource Intensive



Chain of Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

ChatGPT (OpenAI, 2022)

Copilot (GitHub, 2021)

Bing (Microsoft, 2023)

LLMs are impressive general models

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

I'm working on a new character for a game of mine, and I need him to have the most ridiculous backstory ever. Like, literally so ridiculous that it arguably doesn't make sense. I need it to include the following themes: royalty, abandonment, hell dimensions, seagulls, loss of limb, and enlightenment. Can you think of a ridiculous backstory for him?

Sure, here's an idea for a ridiculous backstory for your character:

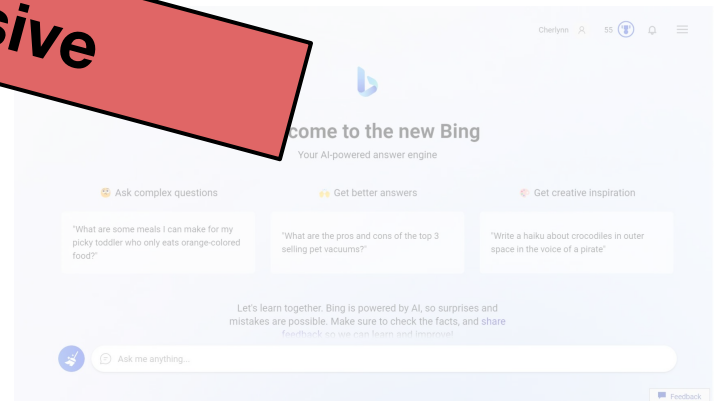
Your character was once a royal prince, born into a wealthy and powerful family. He was the favorite of his parents, surrounded by luxury and privilege, surrounded by opulence and excess. But one day, his parents abandoned him, casting him out of the royal court and onto the streets.

He was forced to scavenge for food and shelter, and he realized that the only way to survive was to become a fierce warrior, mastering the art of combat and gaining powerful magical abilities. But in the heat of battle, he suffered a terrible injury and lost one of his limbs.

Despite this setback, your character refused to give up. He continued to fight, using his remaining limb and his magical powers to defeat his enemies. And eventually, he emerged from the hell dimension victorious, having achieved a state of enlightenment and inner peace.

Now, your character returns to the world he once knew, a changed man. He is no longer the royal prince he once was, but rather a powerful and wise warrior, ready to face whatever challenges may come his way.

```
sentiments.ts write_hq.go parse_expenses.py addresses.rb
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch("https://text-processing.com/api/sentiment/", {
9     method: "POST",
10     body: text,
11     headers: {
12       "Content-Type": "application/x-www-form-urlencoded",
13     },
14   });
15   const json = await response.json();
16   return json.sentiment === "pos";
17 }
```



Chain of Thought Prompting Elicits Reasoning in Large Language Models (Wei et al., 2022)

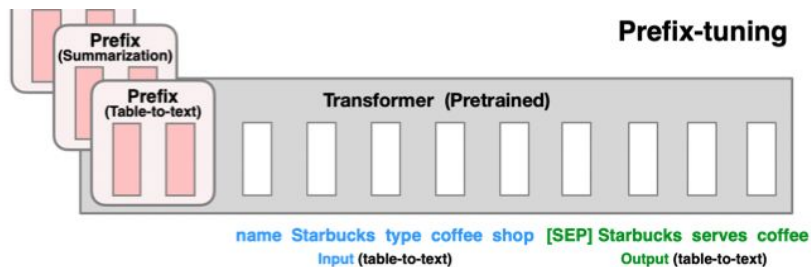
Large Language Models are Zero-Shot Reasoners (Kojima et al., 2022)

ChatGPT (OpenAI, 2022)

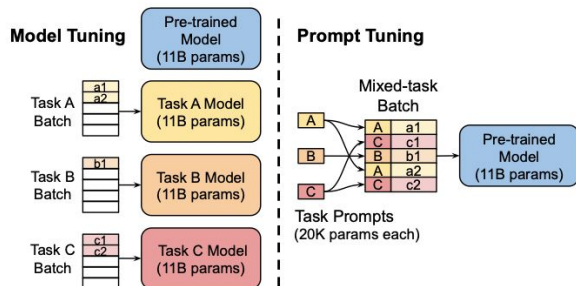
Copilot (GitHub, 2021)

Bing (Microsoft, 2023)

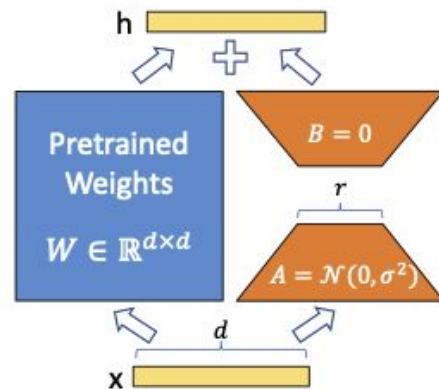
Parameter Efficient Adaptation



Prefix Tuning: Learns a prefix embedding (for each layer) to adapt to new tasks. ~99.9% of the model kept frozen.

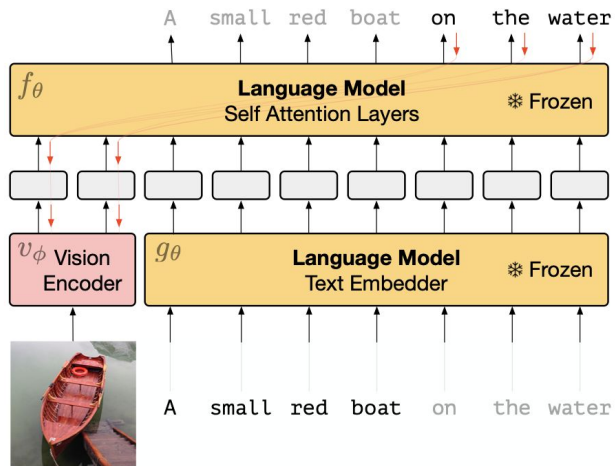


Prompt Tuning: Similar idea to prefix-tuning, but learns just a single prefix for input embeddings.

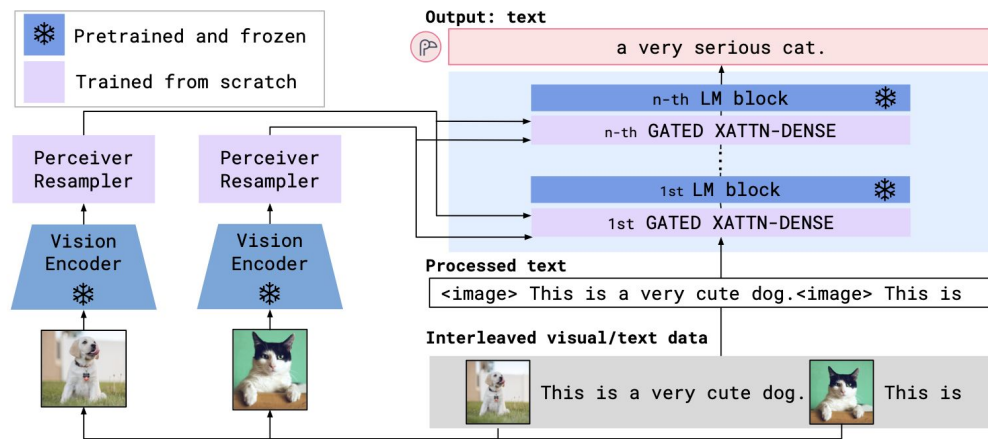


Low-Rank Adaptation: Injects trainable rank decomposition matrices into each Transformer layer of a pretrained model.

Adapting Text-Only LLMs for Multi-Modal Tasks

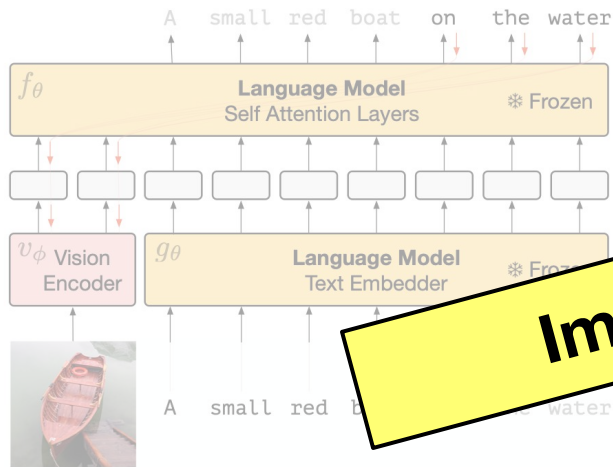


Prefix tuning for adapting LLMs to image captioning.
~95% of the model kept frozen. Capable of
compelling few-shot multi-modal reasoning.



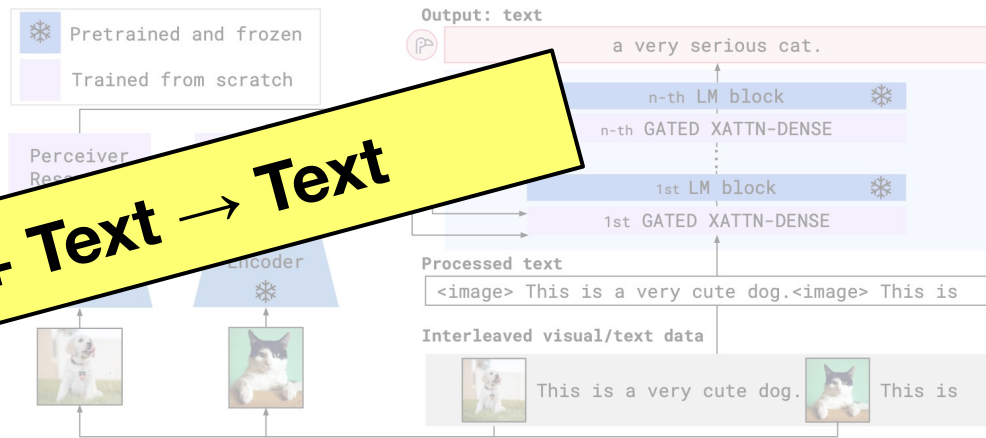
Finetunes new cross-attention layers on top of a 70B LLM.
Achieves SOTA on many multi-modal tasks.

Adapting Text-Only LLMs for Multi-Modal Tasks



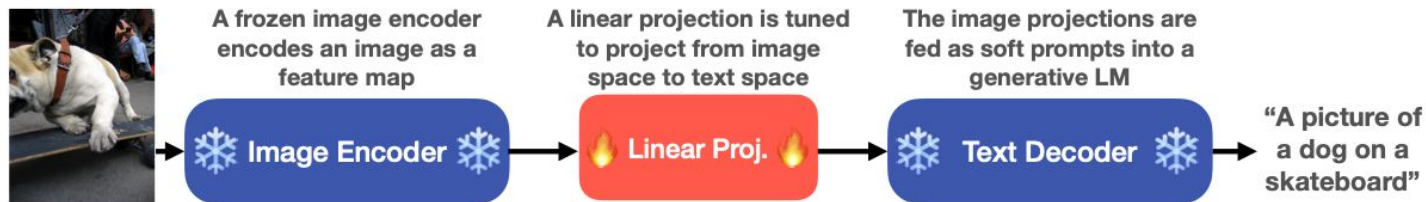
Prefix tuning for adapting LLMs to image captioning.
~95% of the model kept frozen. Capable of
compelling few-shot multi-modal reasoning.

Image + Text → Text



Finetunes new cross-attention layers on top of a 70B LLM.
Achieves SOTA on many multi-modal tasks.

How visually grounded are text-only LLMs?



Merullo et al. showed that pretrained text-only LMs and pretrained visual encoders produce functionally equivalent representations up to a linear mapping.

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



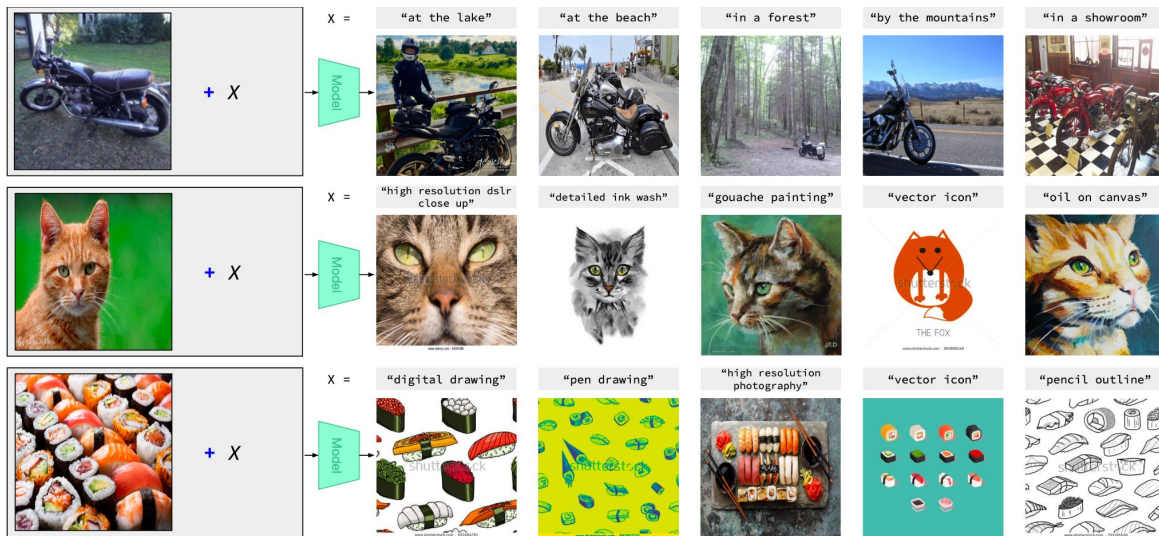
Frozen Retrieval Over Multimodal Data for Autoregressive Generation

jykoh.com/fromage

FROMAGe

Frozen Retrieval Over Multimodal Data for Autoregressive Generation

jykoh.com/fromage

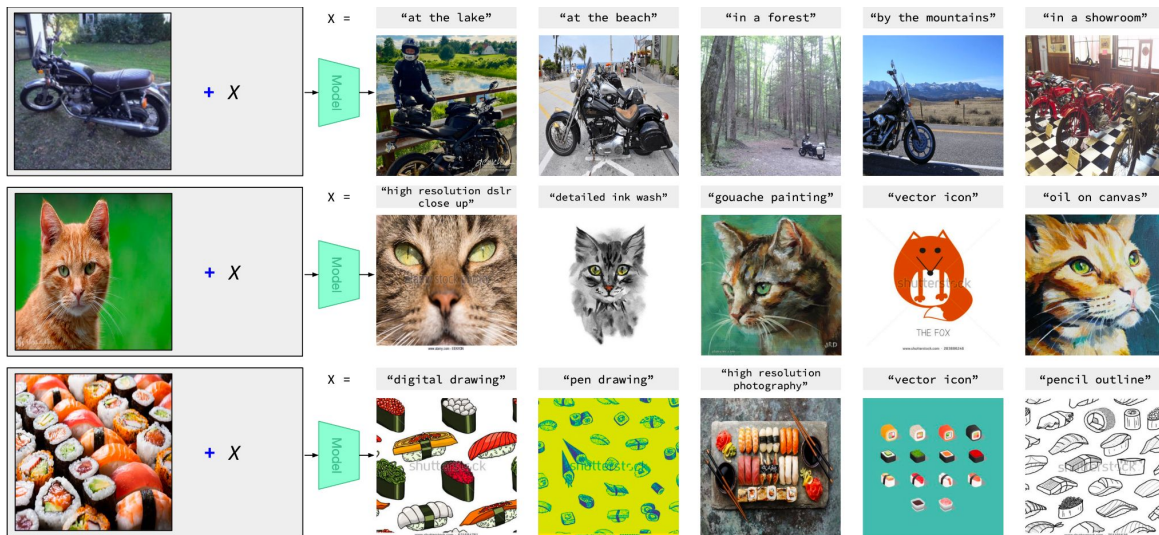


Concept composition. Our model is capable of retrieving relevant images conditioned on multi-modal context inputs.



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



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

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

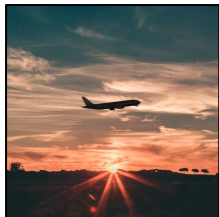


Image #1

silhouette
of a plane
against
the sunset

Caption #1

Image Captioning

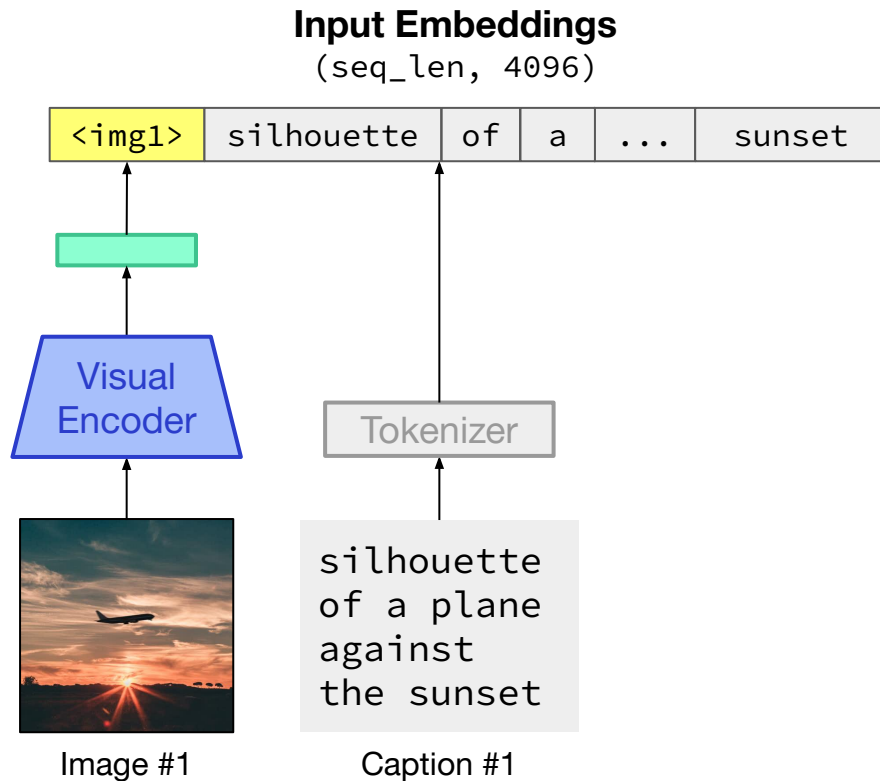
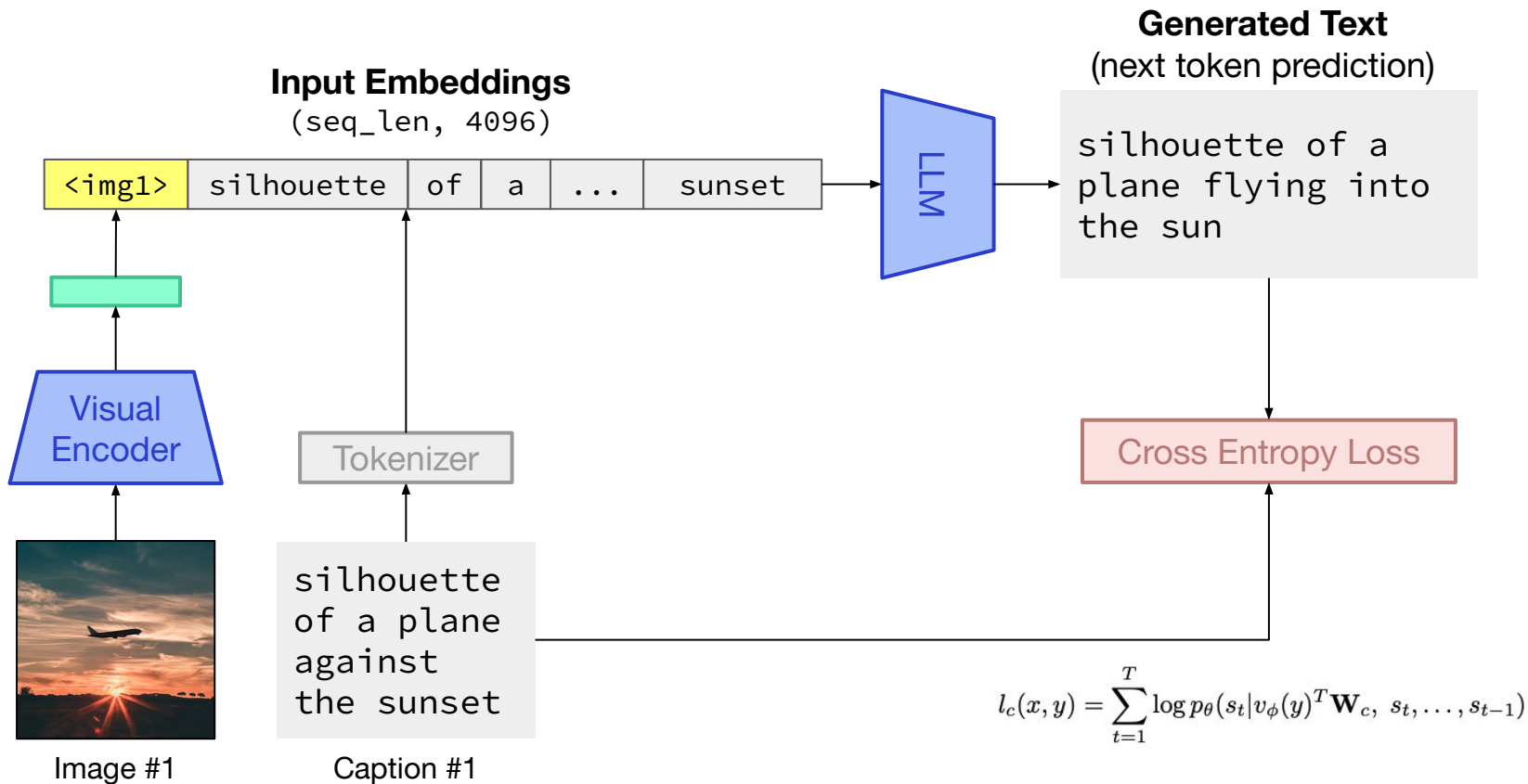


Image Captioning



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

Bootstrapping LLMs for Image-Text Retrieval

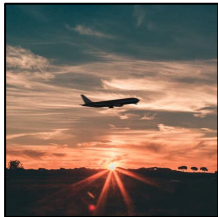
- SOTA image-text retrieval models (CLIP, ALIGN) usually use encoder-based language models

Bootstrapping LLMs for Image-Text Retrieval

- SOTA image-text retrieval models (CLIP, ALIGN) usually use encoder-based language models
- **How do we adapt an autoregressive language model for this?**
 - Learn a special [RET] token for retrieving images
 - Train the model to learn when to generate [RET]
 - Improves retrieval by ~37% over having no dedicated [RET] token

silhouette of
a plane
against the
sunset [RET]

Input Caption



Input Image

Image-Text Retrieval

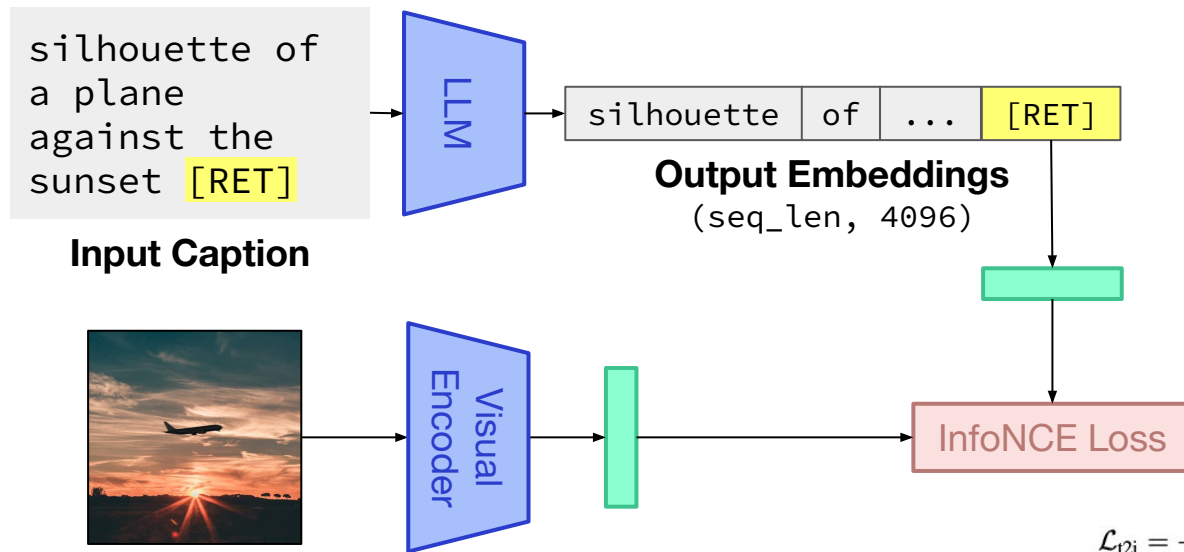
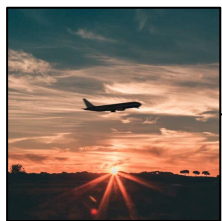
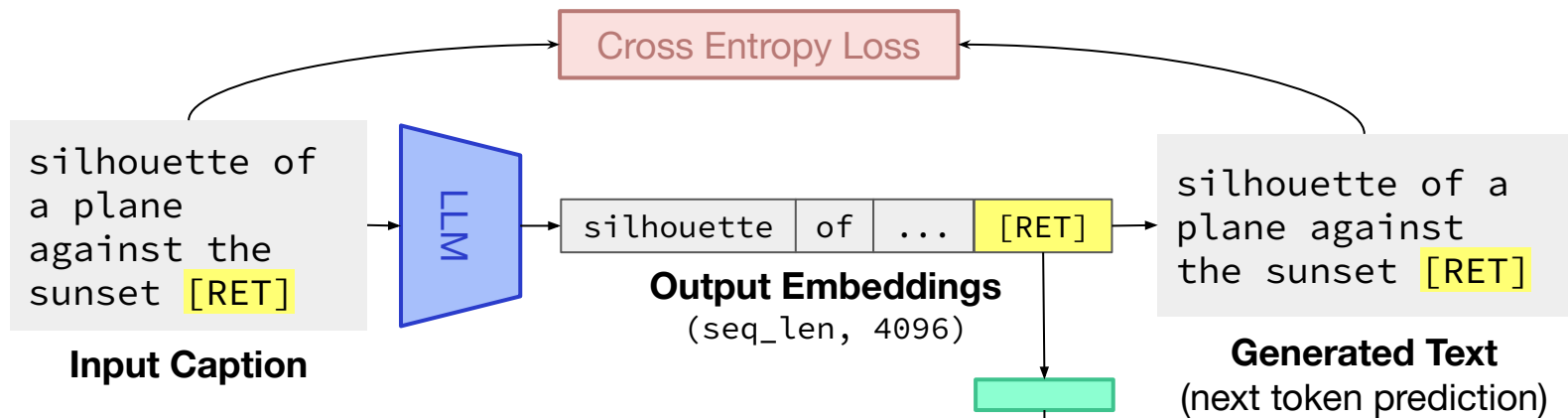


Image-Text Retrieval

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

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



Input Image

Image-Text Retrieval

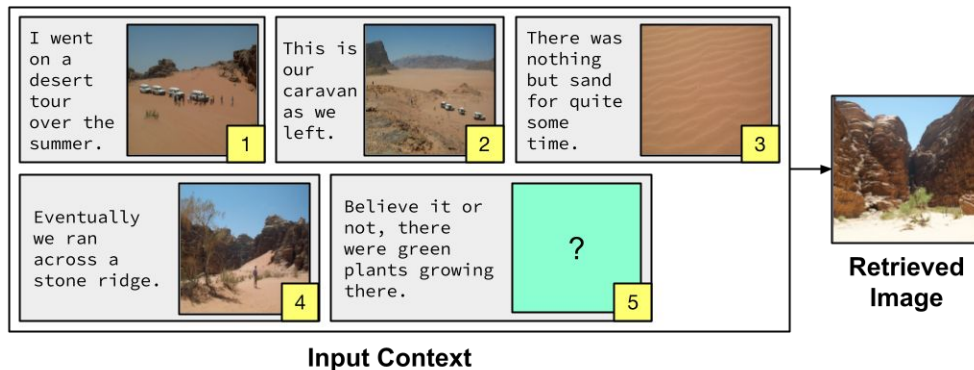
$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N} \sum_{i=1}^N \left(\log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(x_i, y_j)/\tau)} \right)$$

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

Quantitative Evaluations

1) Contextual image retrieval

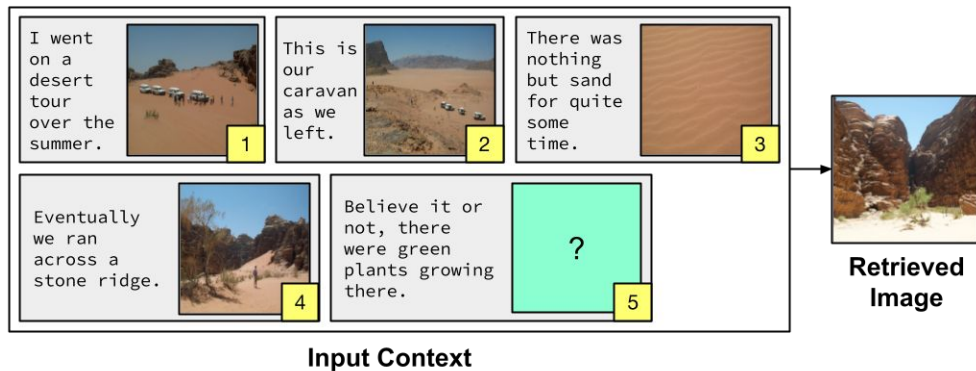
- Given a Visual Story, retrieve the correct image
- FROMAGE is more sensitive to context
- CLIP gets worse with more context



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



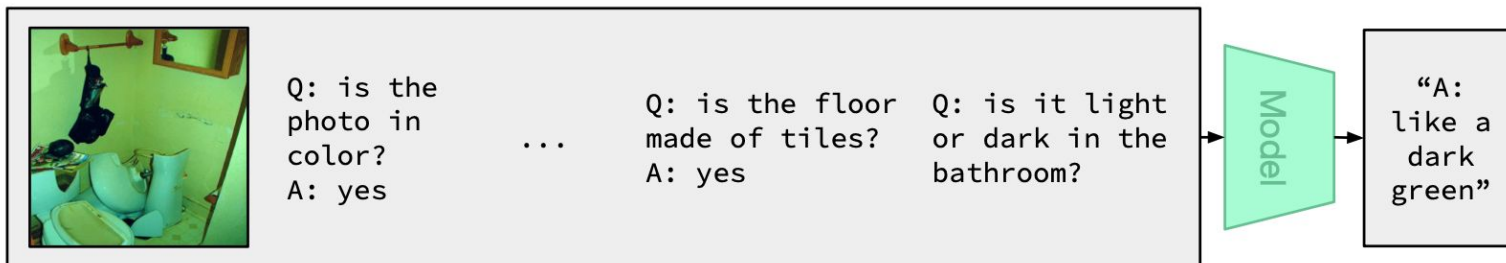
Model	Inputs	R@1	R@5	R@10
CLIP ViT-L/14	1 caption	11.9	25.5	32.2
FROMAGe		9.0	21.1	28.7
CLIP ViT-L/14	5 captions	5.9	19.5	28.0
FROMAGe		10.4	23.8	31.7
CLIP ViT-L/14	5 captions, 4 images	Incapable		
CLIP ViT-L/14 [†]	5 captions	8.8	22.3	29.8
FROMAGe [†]	5 captions	11.6	24.7	32.8
FROMAGe [†]	5 captions, 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.

Quantitative Evaluations

2) Visual Dialogue

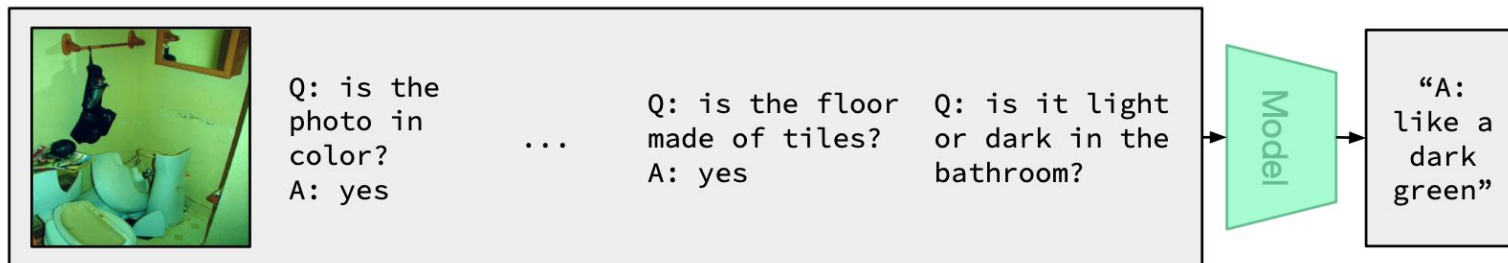
- IT2T: Answer a question about the image given past dialogue discussing it



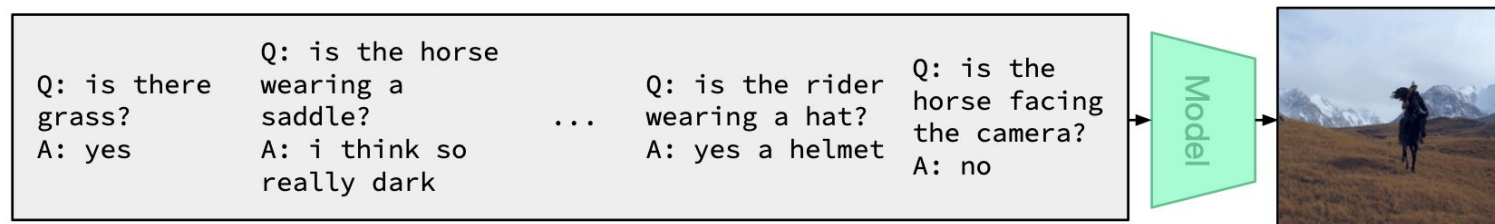
Quantitative Evaluations

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



Quantitative Evaluations

2) Visual Dialogue

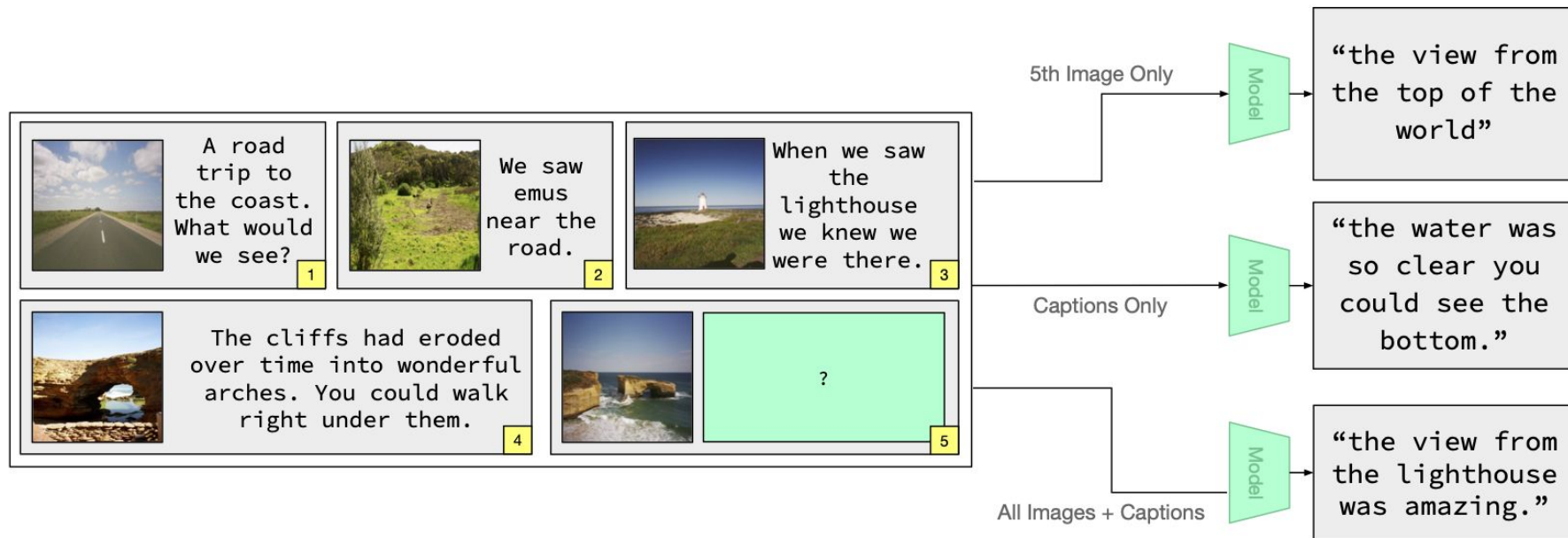
- **IT2T**: Answer a question about the image given past dialogue discussing it
- **T2I**: Retrieve the correct image given a series of dialogue about it

Model	Trainable Params	Dataset Size	IT2T					T2I		
			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.

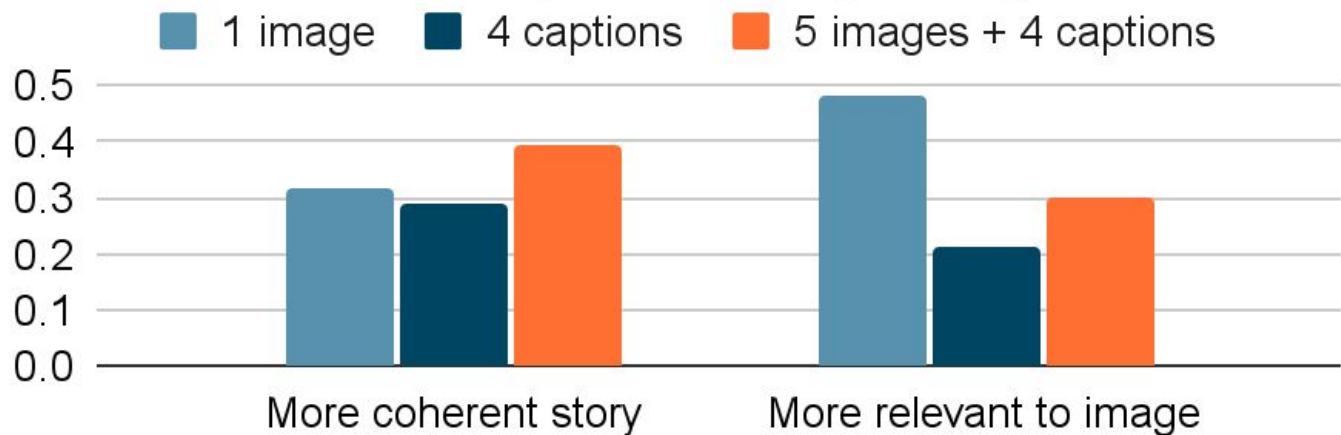
Human Evaluations (Text Generation)

Does additional **multi-modal context** help in generating **good stories**?



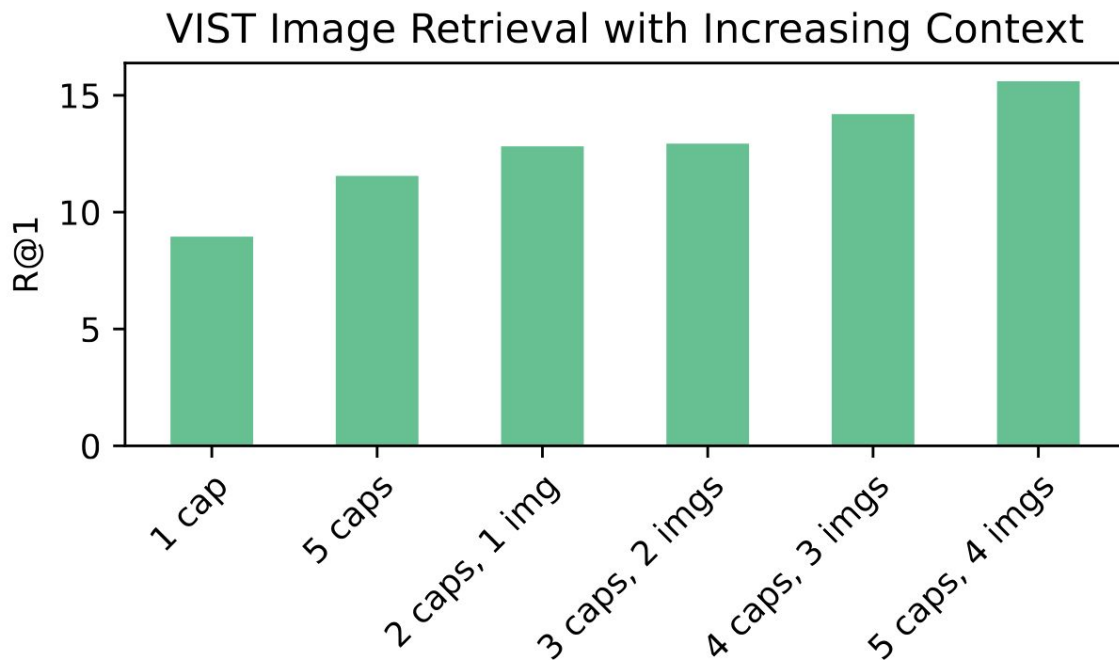
Human Evaluations (Text Generation)

Human Preference (Visual Storytelling)

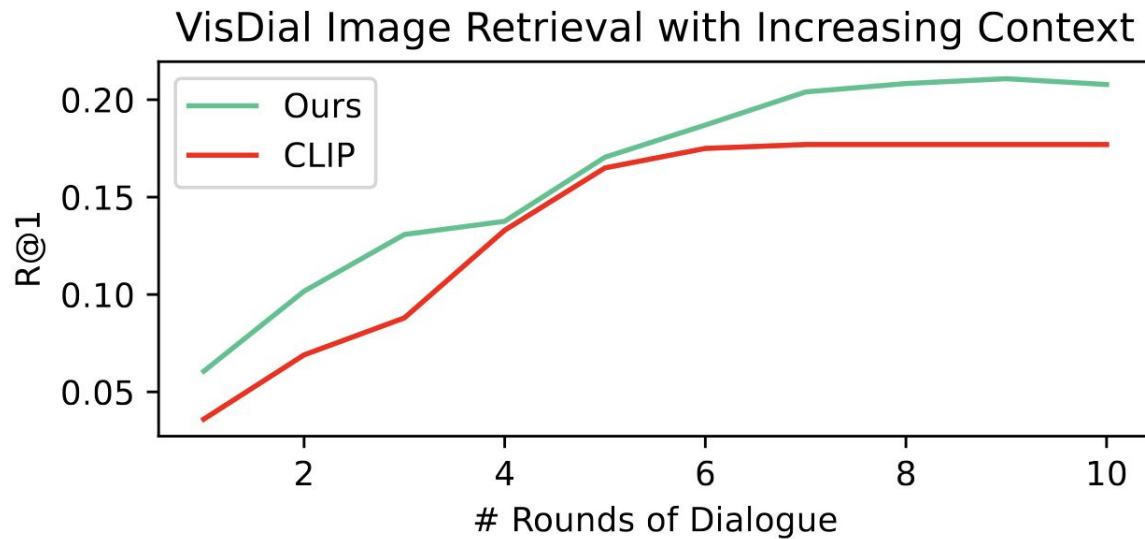


The Effect of Context

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



The Effect of Context



Freezing is Essential

- Finetuning the LLM leads to lower training and validation loss
- But the model loses its in-context learning ability!

	<i>Frozen</i>	Finetuned
VIST R@1	12.8 →	6.2
VisDial R@1	14.6 →	1.0

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.



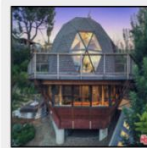
Yes, what about a pencil drawing of one?



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.

Longer Visual Prefix → Improved Dialogue Generation

Melting Earth Ice Cream



What is the meaning of this?

The melting ice cream is a metaphor for the melting of the polar ice caps.

What does it provoke?

It provokes a sense of urgency to save the planet.

How does it do that?

It shows the world is in danger and we need to act now.

Is it effective?

Yes, it is effective because it is a very powerful image.

Who is it targeted at?

It is targeted at everyone.

What will people do after seeing this?

They will be motivated to act.

Show me some examples of things they can do

They can sign petitions, write letters to their representatives, and donate to environmental organizations.

Come up with a tagline for this image

Save the planet.

Octopus Country Political Cartoon



What is this cartoon about?

This cartoon is about the political and economic power of the country.

How is the country portrayed?

The country is portrayed as a giant octopus with tentacles reaching out to the world.

What can you infer from this?

The cartoon implies that the country is a powerful country that is trying to control the world.

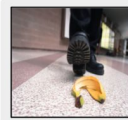
What does it provoke from its target audience?

The cartoon provokes the audience to think about the country and its power.

Where can I put this to good use?

You can put this to good use in your classroom to teach about the country and its power.

Banana Peel



What is about to happen?

A person is about to fall down.

How will he do that?

He will trip over a banana peel.

What should he do?

He should pick up the banana and throw it away.

Where can he do that?

In the garbage can.

Sofa Recommendations



What is this?

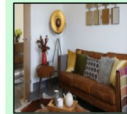
This is a sofa that is made from a single piece of leather.

Where would this look good in?

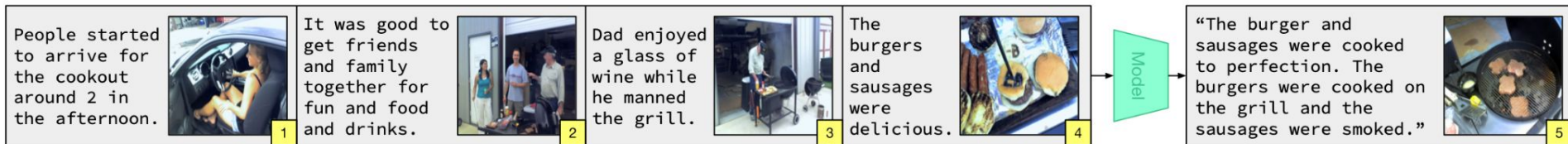
This sofa would look great in a living room or a bedroom.



leather sofa with a coffee table and a lamp.

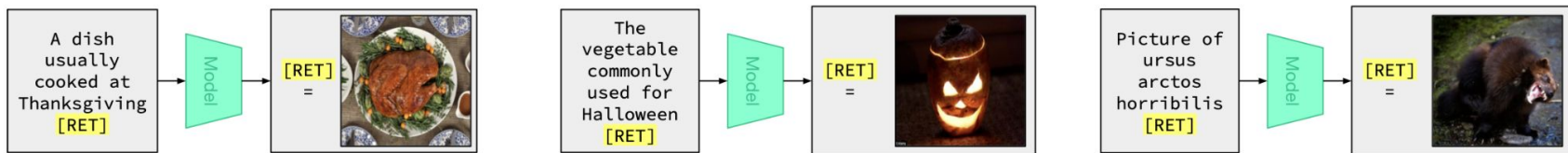


In-Context Learning and Other Abilities



Visual Storytelling

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



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.



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.

Future Work

- **Train on more diverse data**
 - CC3M is small by modern standards – we would get a lot more from training on LAION
- **Generate images from scratch rather than retrieve**
- **Train more sophisticated image-text mappings**
 - Adapters, cross-attention layers, LoRA
- **Apply to even larger LLMs and stronger visual models**

Try the model!

huggingface.co/spaces/jykoh/fromage

Spaces: jykoh **fromage** like 6 Running on A100 Open logs

App Files and versions Community Settings

Linked Models

FROMage

This is the official Gradio demo for the FROMage model, a model that can process arbitrarily interleaved image and text inputs, and produce image and text outputs.

Paper: [Grounding Language Models to Images for Multimodal Generation](#)

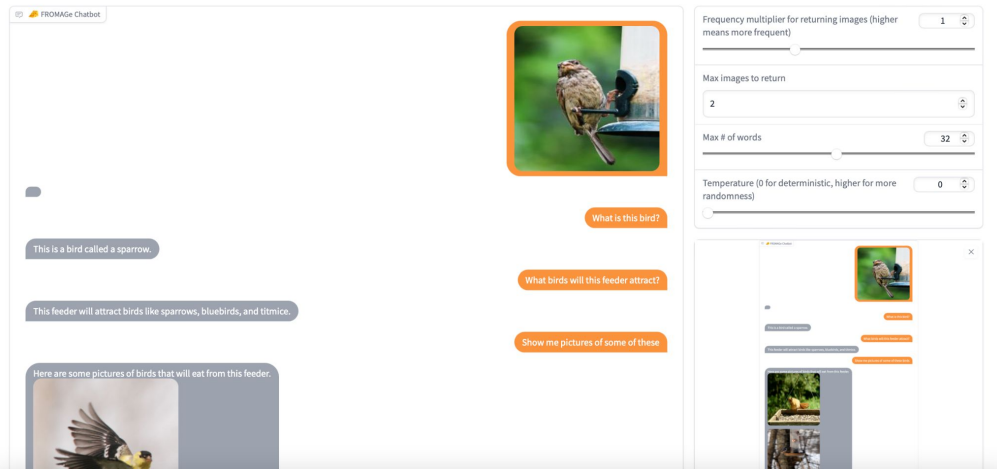
Project Website: [FROMage Website](#)

Code and Models: [GitHub](#)

Tips:

- Start by inputting either image or text prompts (or both) and chat with FROMage to get image-and-text replies.
- Tweak the level of sensitivity to images and text using the parameters on the right.
- Check out cool conversations in the examples or community tab for inspiration and share your own!
- For faster inference without waiting in queue, you may duplicate the space and use your own GPU: [Duplicate Space](#)

FROMage Chatbot



FROMage Chatbot

What is the difference between a biscuit in the United States and the United Kingdom?

In the United States, a biscuit is a small, round, baked, sweet, and sometimes savory food.

Show me what that looks like

In the United States



What about the ones in the United Kingdom?

In the United Kingdom



Thanks!

jkoh@cmu.edu

jkoh.com/fromage