Generating Images with Multimodal Language Models

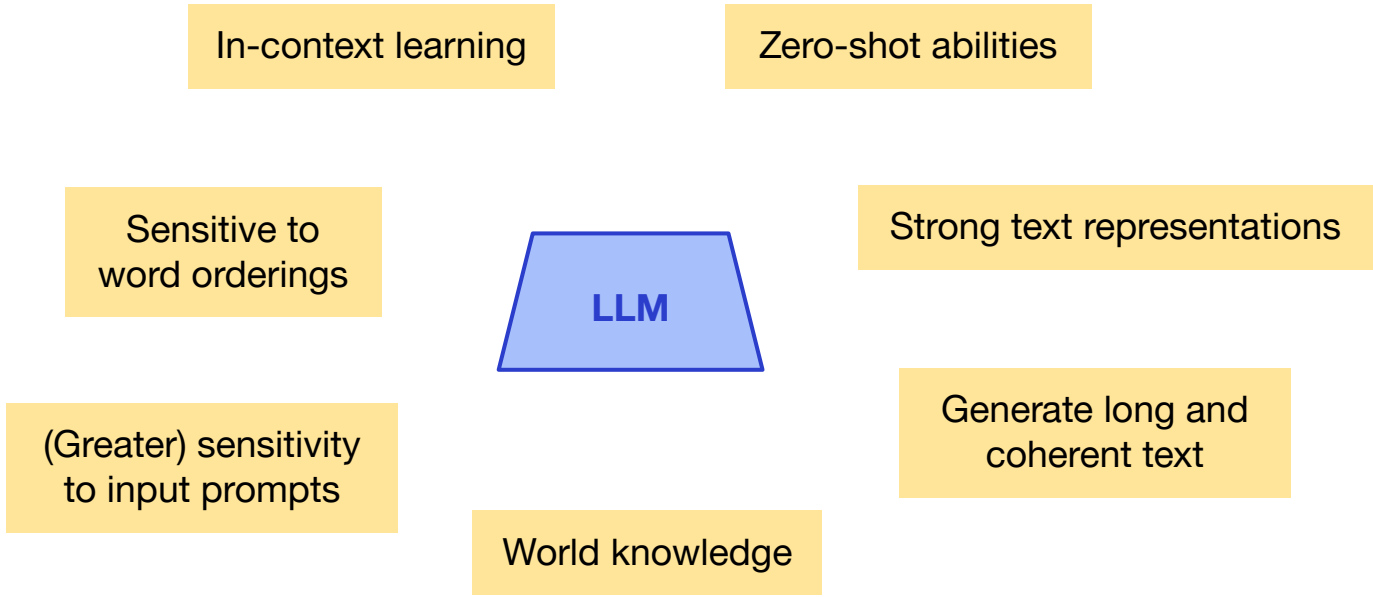
jykoh.com/gill

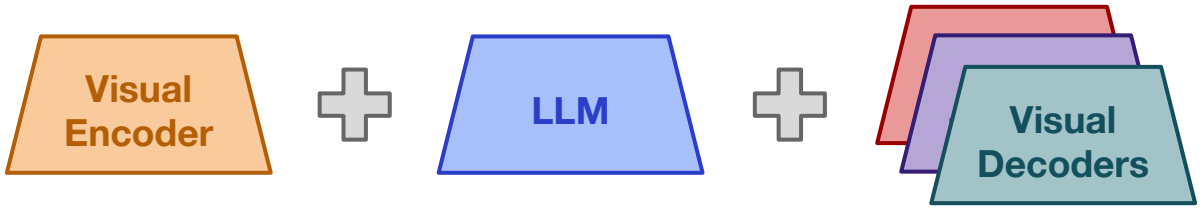
Jing Yu Koh, Daniel Fried, Ruslan Salakhutdinov

6 September 2023

Cohere for AI

Carnegie Mellon University

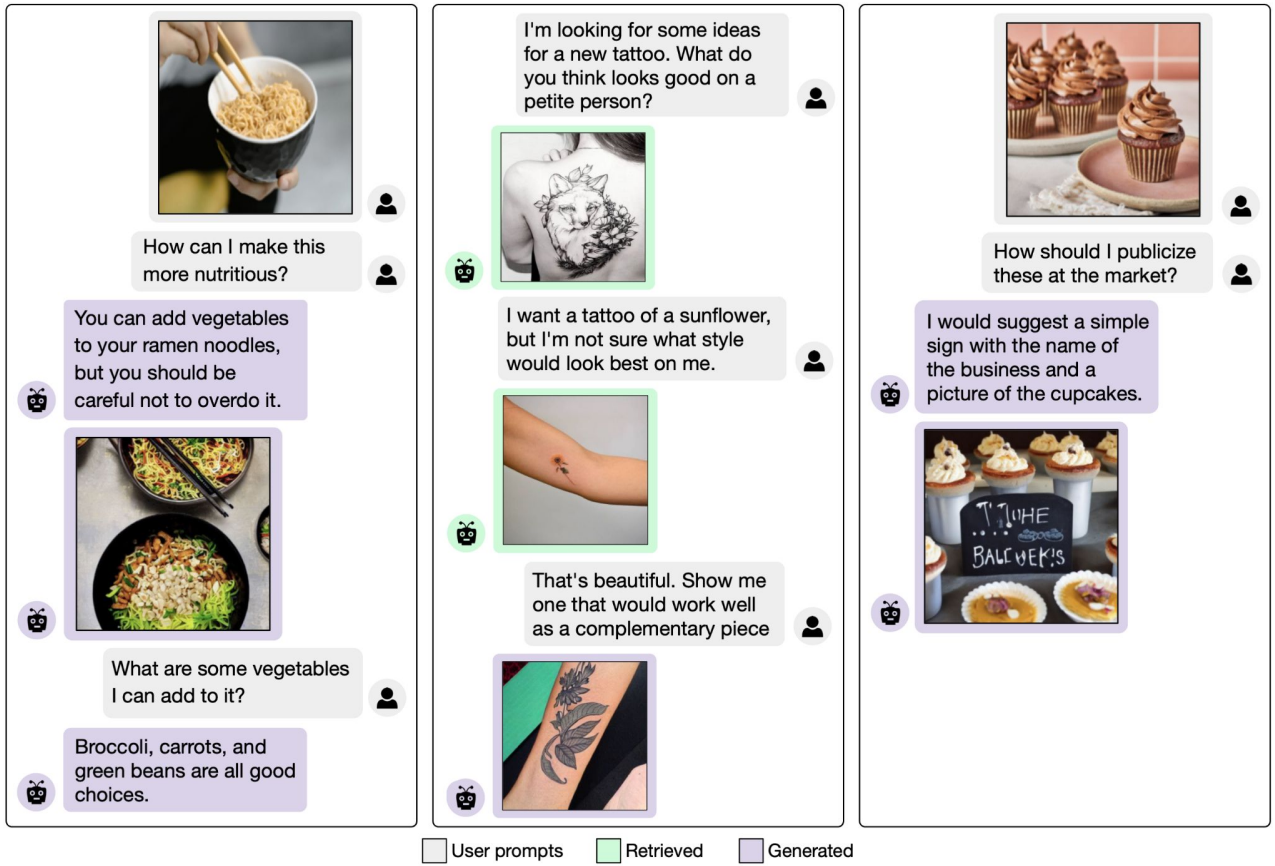

In-context learning
Zero-shot abilities
Sensitive to word orderings
LLM
Strong text representations
(Greater) sensitivity to input prompts
Generate long and coherent text
World knowledge



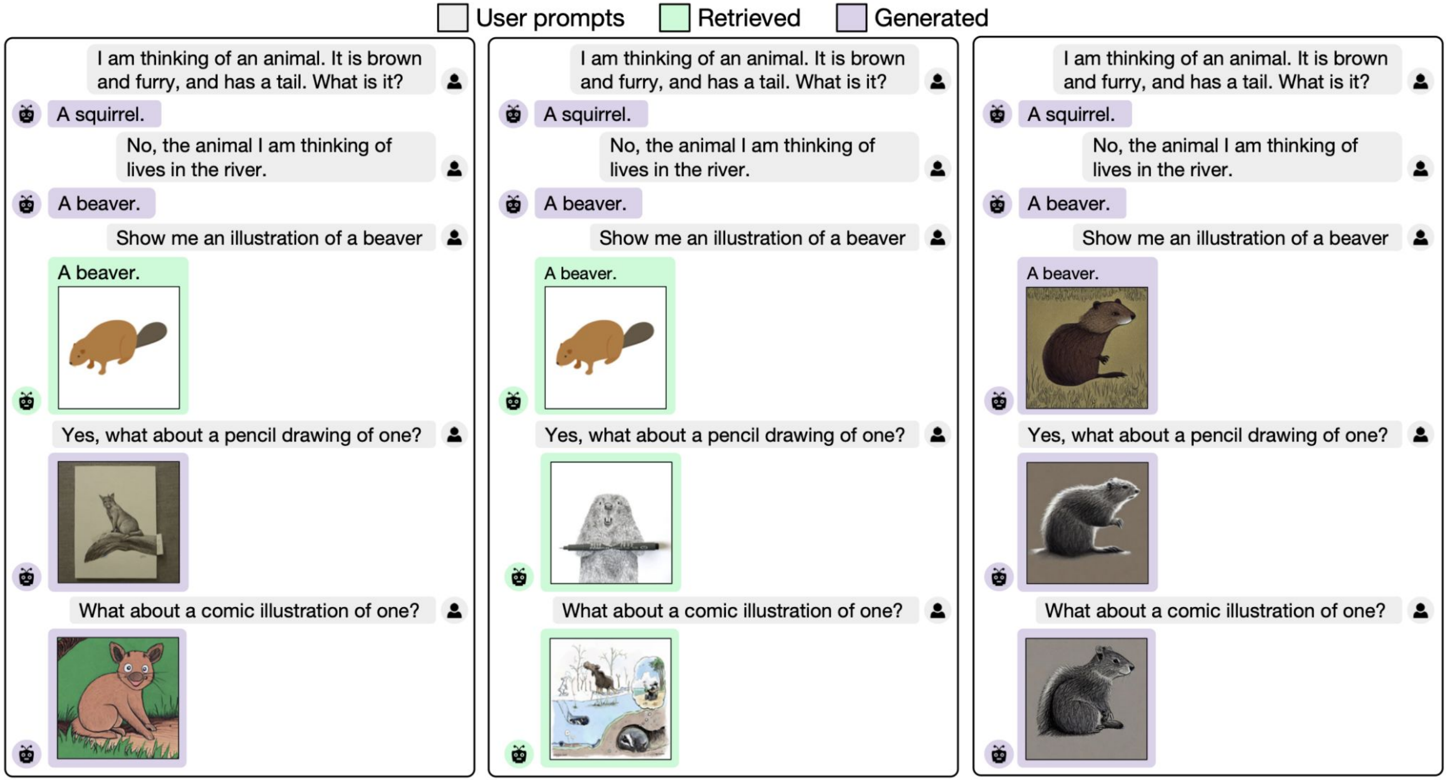
Can we *ground* text-only LLMs to pretrained visual encoders and decoders?

GILL

Generating Images with Large Language Models



GILL: A More General Multimodal LM



Ours
Retrieval and generation multimodal LM. Decides when to retrieve or generate.

FROMAGe
Retrieval only multimodal language model. Image outputs limited to the candidate retrieval set.

Stable Diffusion
Generation only text-to-image model. Less sensitive to longer text inputs (such as dialogue).

🐟 GILL: A More General Multimodal LM

- **Frozen** (Tsimpoukelli et al., 2021)
  Flamingo (Alayrac et al., 2022)
  BLIP-2 (Li et al., 2023)
  - Process image + text, generate text only
- **FROMAGe** (Koh et al., 2023)
  - Process image + text, generate text + retrieve images
- **GILL** (this work)
  - Process image + text, generate text + retrieve images + generate images
  - Decides whether to retrieve images or generate from scratch
  - Resource efficient: trained on 2 GPUs for 2 days

🐟 GILL

Generating Images with Large Language Models

- **Capable of retrieving images, generating images, and generating text**
  - Can condition on arbitrarily interleaved image + text inputs
  - Generate text, generate images, and retrieve images as part of the output
- **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 7B LLM, the CLIP encoder, and the Stable Diffusion image generator
  - Likely benefits from using larger and stronger LLMs in the future
  - Can be applied with other visual models (e.g., OCR) to introduce new abilities

Learning to *Process* Images

Image #1

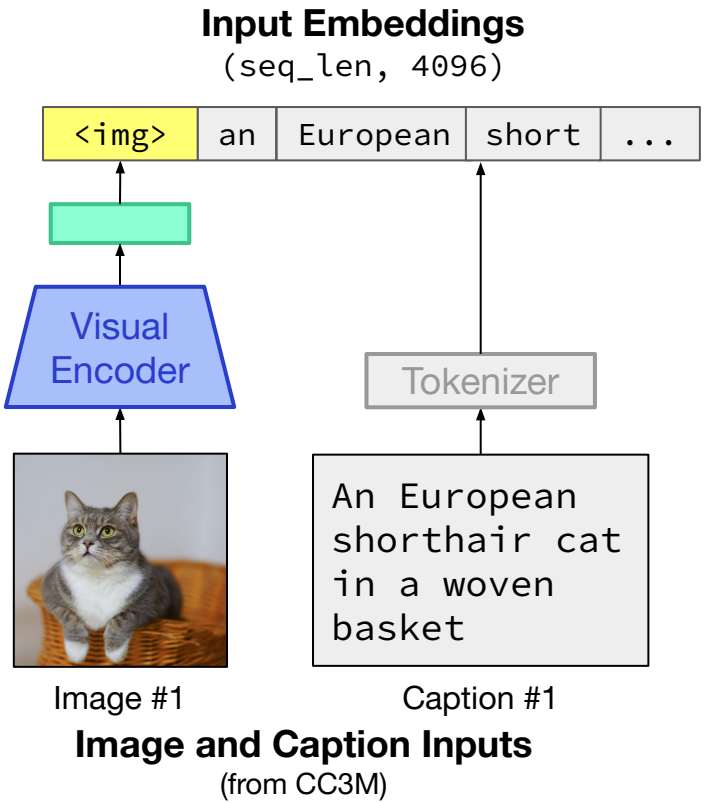
An European shorthair cat in a woven basket

Caption #1

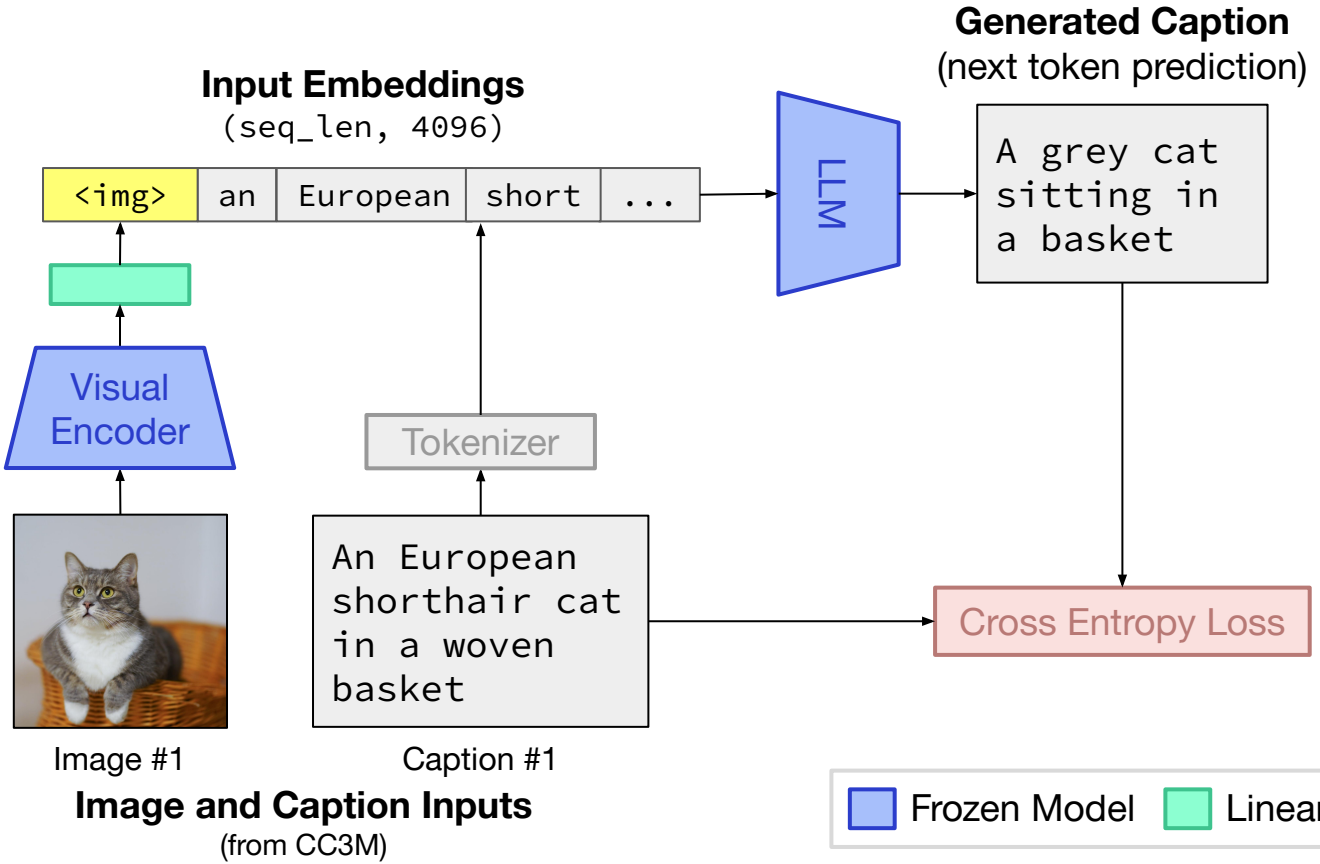
Image and Caption Inputs
(from CC3M)

Frozen Model | Linear Layer | Loss

Learning to *Process* Images



Learning to *Process* Images



Learning to *Produce* Images

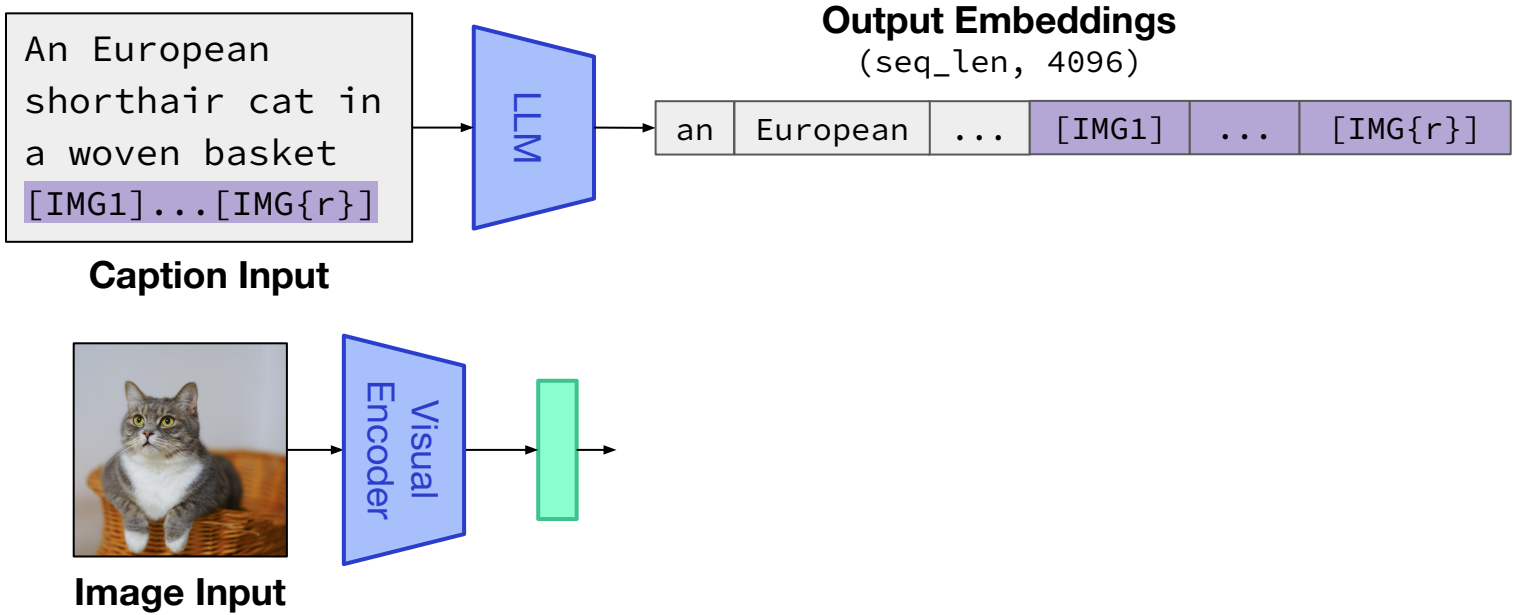
An European shorthair cat in a woven basket [IMG1]...[IMG{r}]

Caption Input

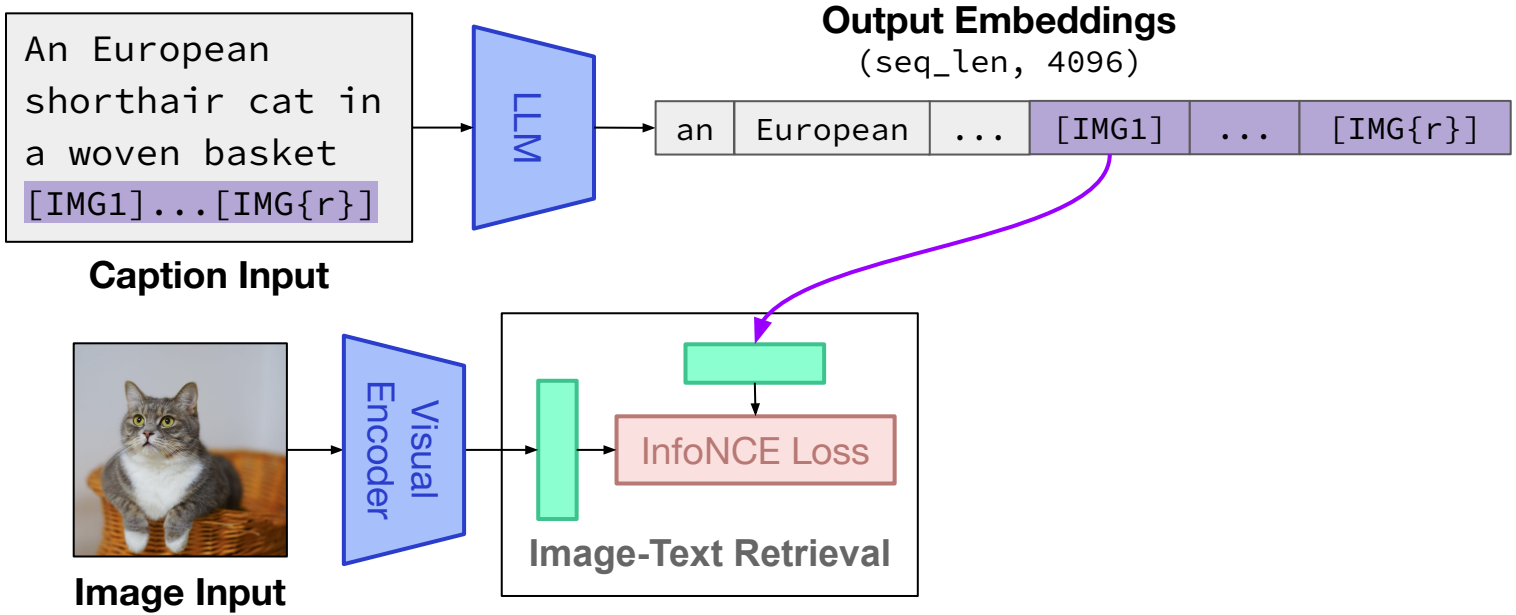
Image Input

Frozen Model | GILLMapper | Linear Layer | Loss

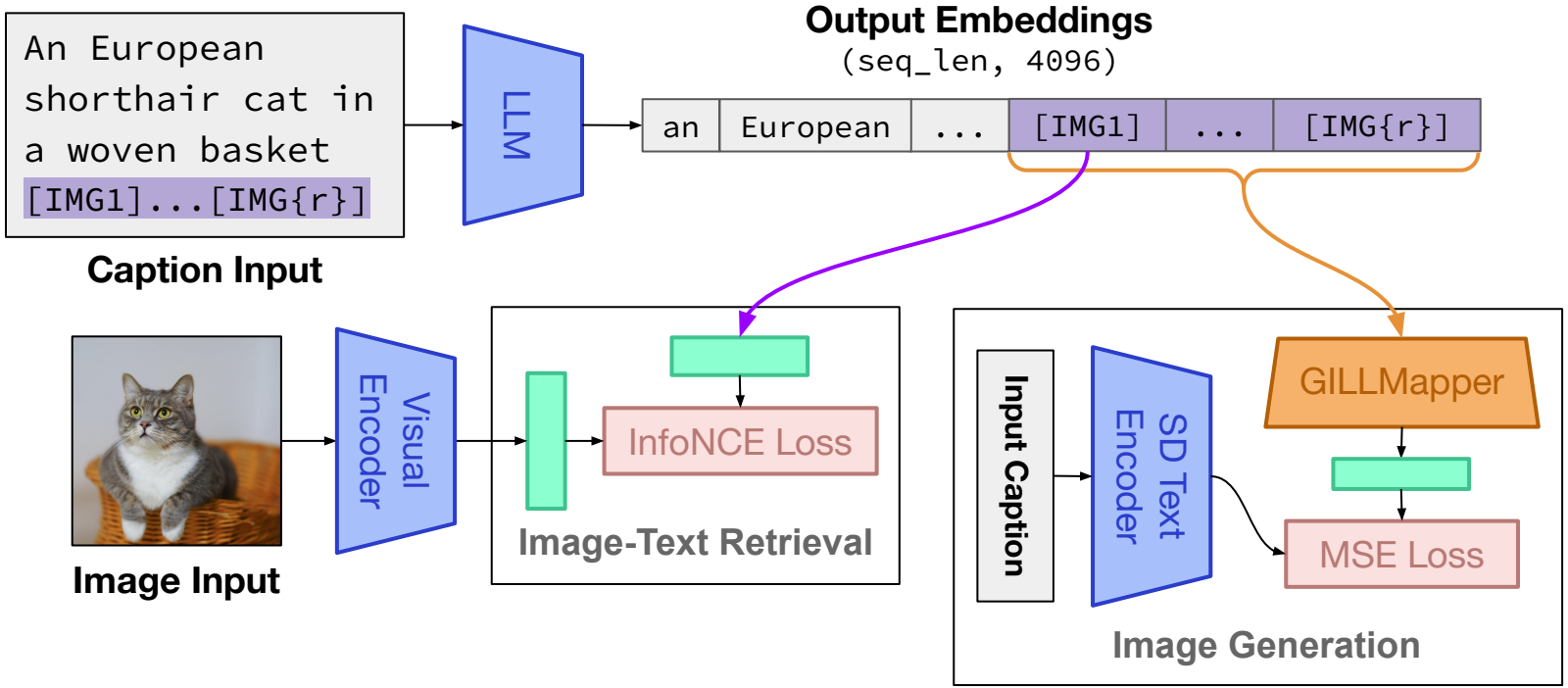
Learning to *Produce* Images



Learning to *Produce* Images

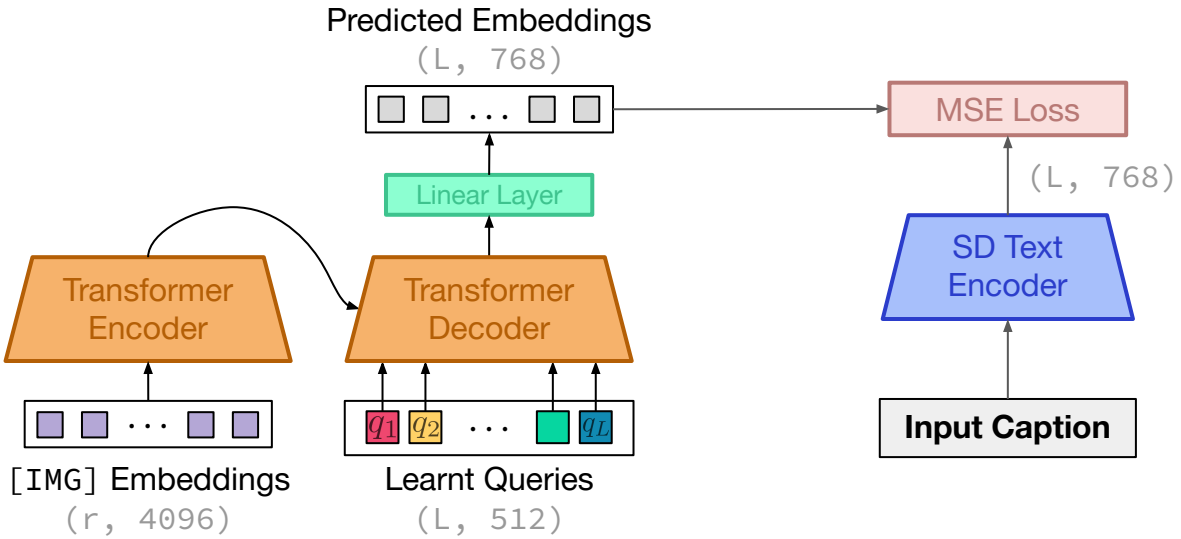


Learning to *Produce* Images

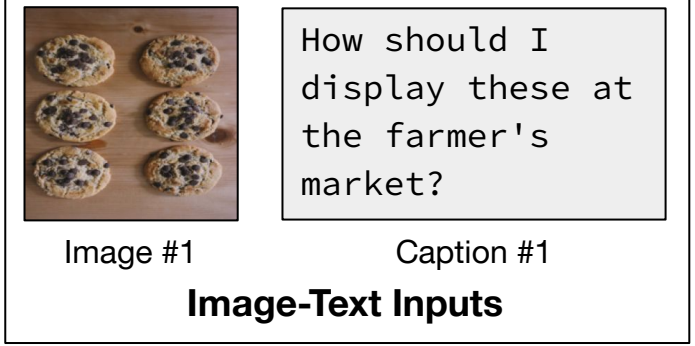


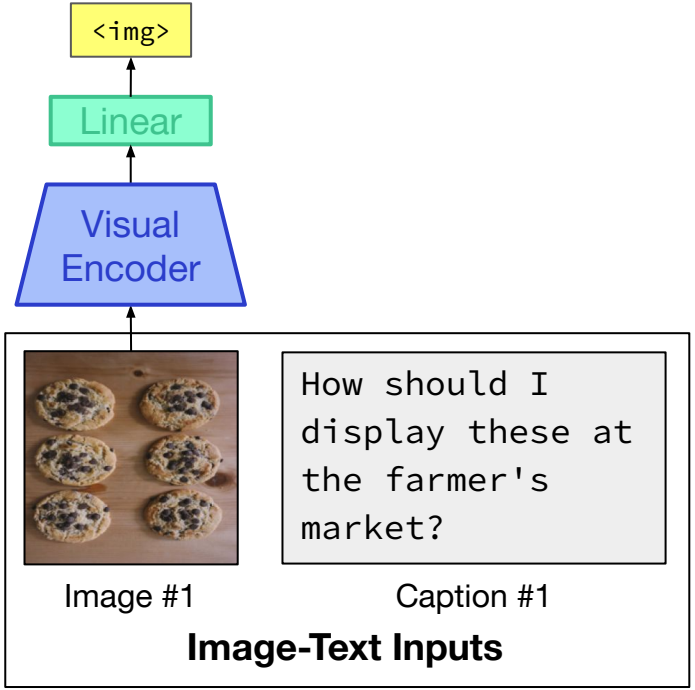
GILLMapper: An Improved LLM-to-Generator Map

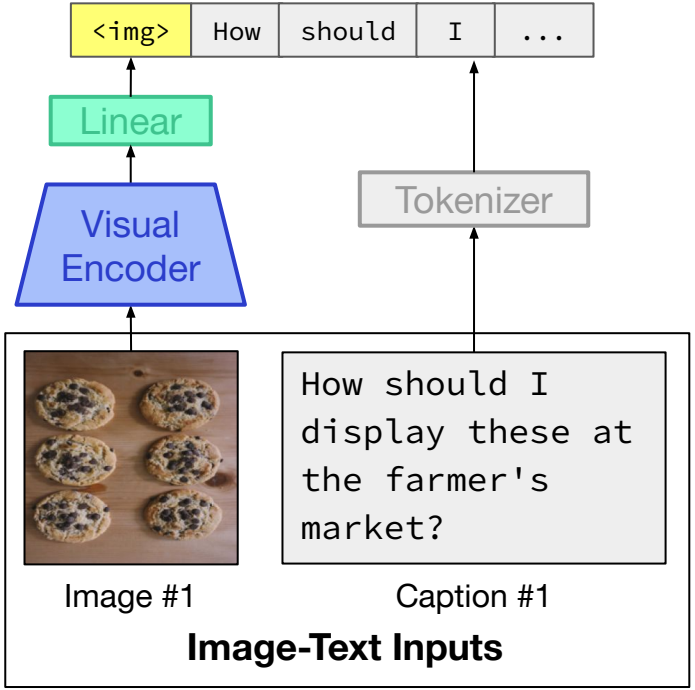
- Previous approaches use linear mappings between LLMs and visual models
- This is insufficient for image generation: decoders require dense information

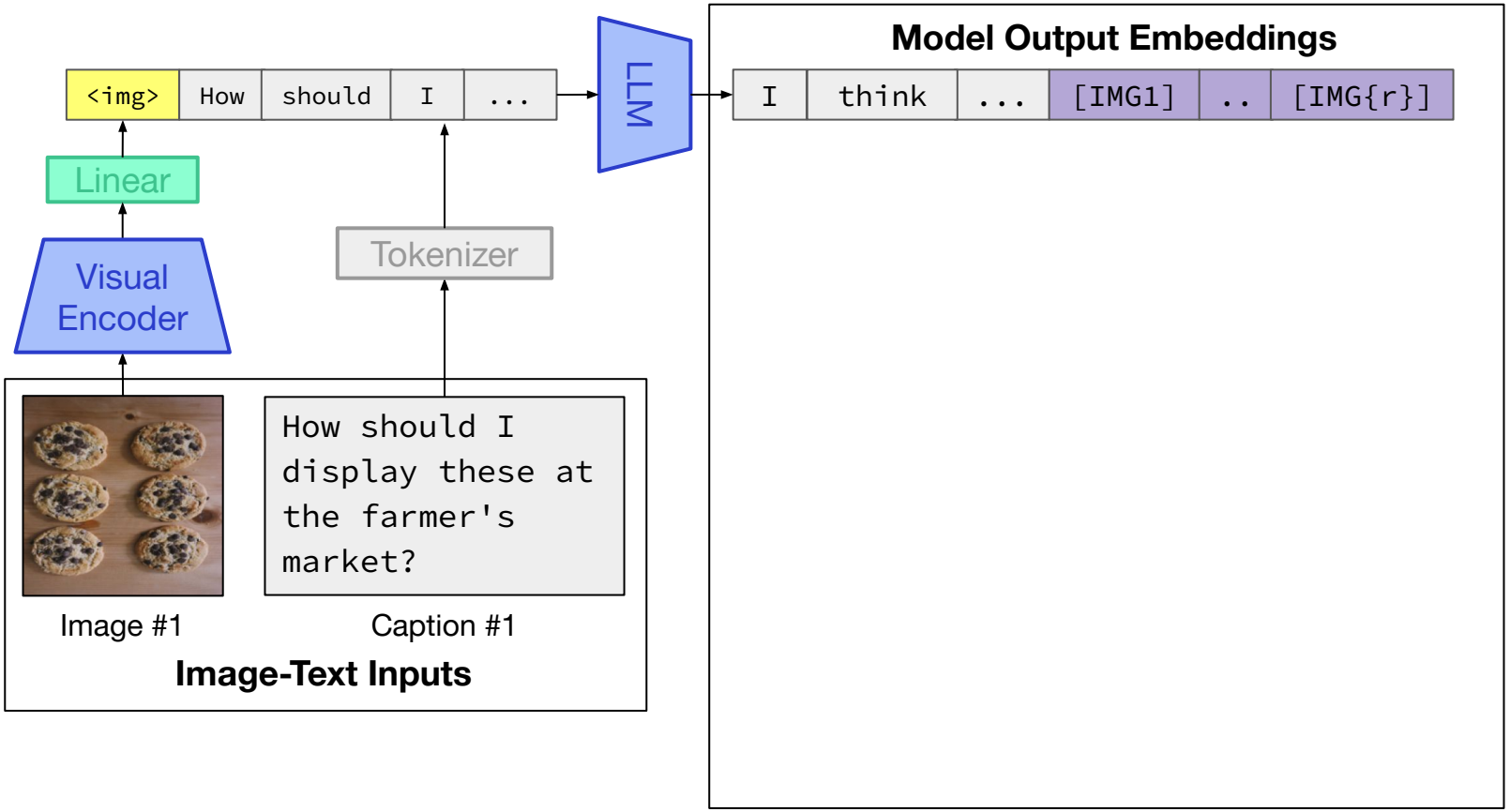


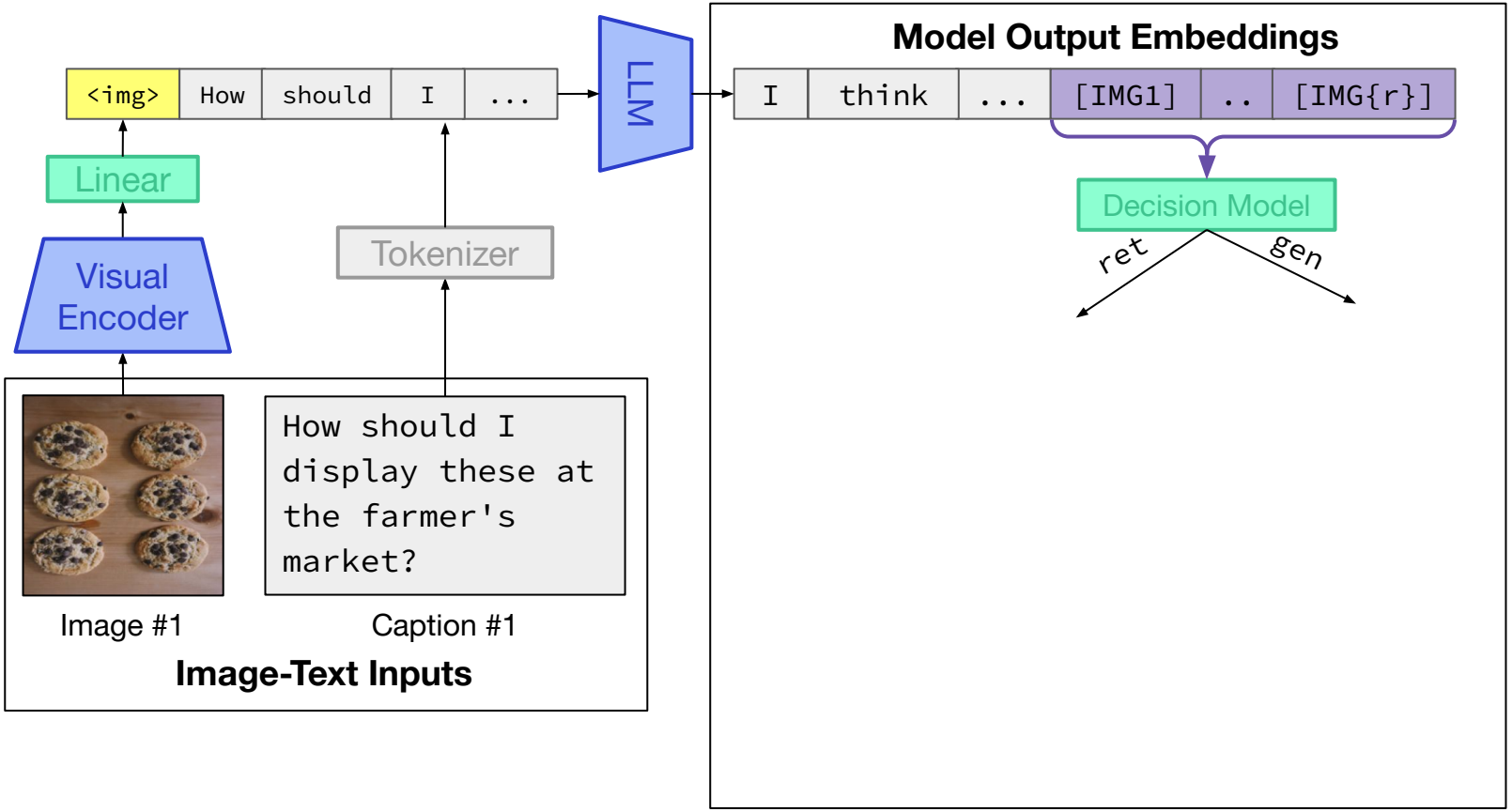
Multimodal Few-Shot Learning with Frozen Language Models (Tsimpoukelli et al., 2021)
Linearly Mapping from Image to Text Space (Merullo et al., 2023)
Grounding Language Models to Images for Multimodal Inputs and Outputs (Koh et al., 2023)

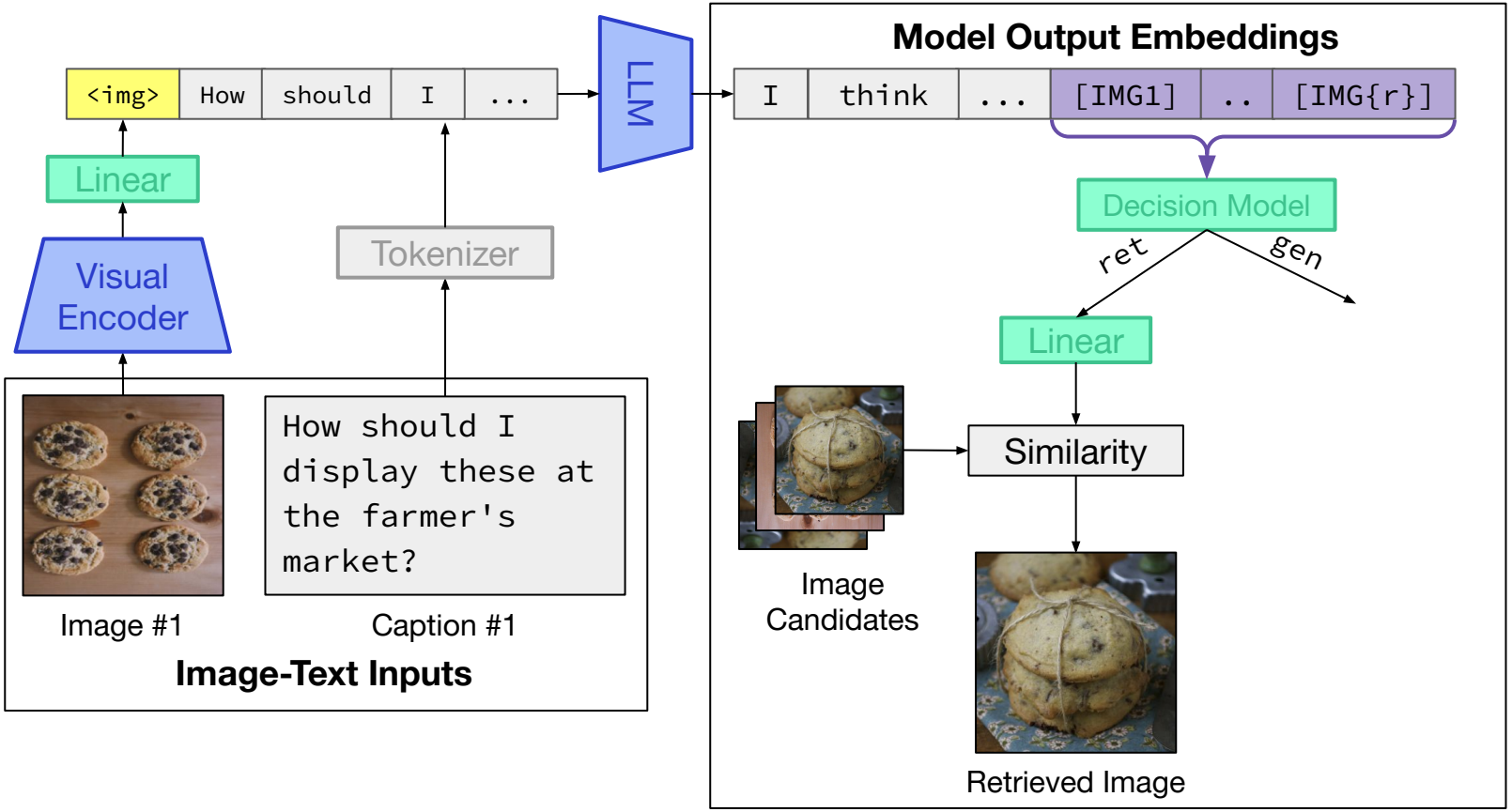

How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs

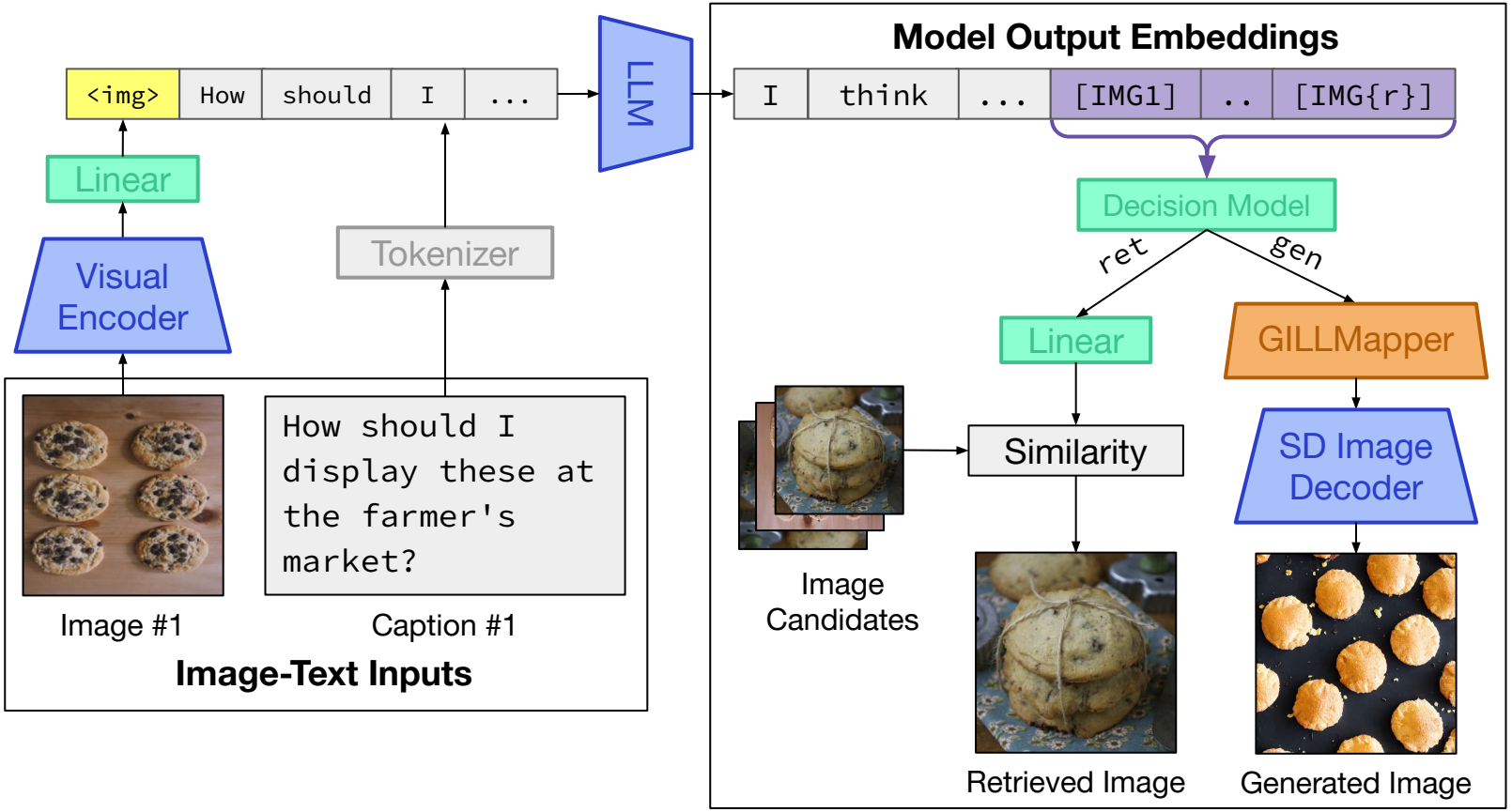

<img>
Linear
Visual Encoder
How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs

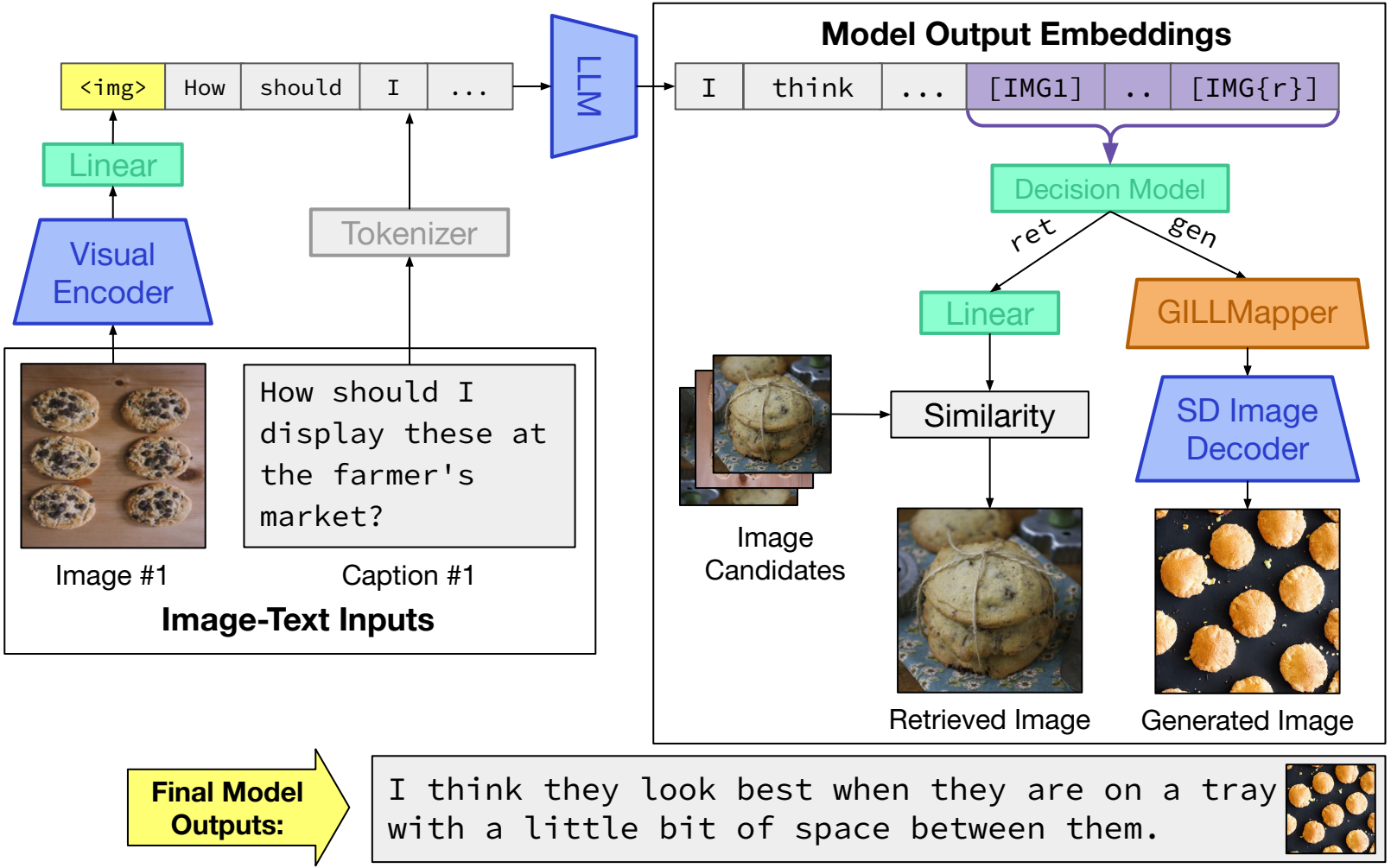

<img>
How
should
I
...
Linear
Visual Encoder
Tokenizer
How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs


<img>
How
should
I
...
LLM
Model Output Embeddings
I
think
...
[IMG1]
..
[IMG{r}]
Linear
Visual Encoder
Tokenizer
How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs


<img>
How
should
I
...
LLM
Model Output Embeddings
I
think
...
[IMG1]
..
[IMG{r}]
Linear
Visual Encoder
Tokenizer
Decision Model
ret
gen
How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs


<img>
How
should
I
...
LLM
Model Output Embeddings
I
think
...
[IMG1]
..
[IMG{r}]
Linear
Visual Encoder
Tokenizer
Decision Model
ret
gen
Linear
How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs
Similarity
Image Candidates
Retrieved Image


<img>
How
should
I
...
LLM
Linear
Visual Encoder
Tokenizer
How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs
Model Output Embeddings
I
think
...
[IMG1]
..
[IMG{r}]
Decision Model
ret
gen
Linear
GILLMapper
Similarity
SD Image Decoder
Image Candidates
Retrieved Image
Generated Image


<img>
How
should
I
...
LLM
Linear
Visual Encoder
Tokenizer
How should I display these at the farmer's market?
Image #1
Caption #1
Image-Text Inputs
Model Output Embeddings
I
think
...
[IMG1]
..
[IMG{r}]
Decision Model
ret
gen
Linear
GILLMapper
Similarity
SD Image Decoder
Image Candidates
Retrieved Image
Generated Image
Final Model Outputs:
I think they look best when they are on a tray with a little bit of space between them.

Evaluation: Contextual Image Generation

- Given a Visual Story, generate a relevant image

Image and Text Inputs

Visual Storytelling (Huang et al., 2016)

Evaluation: Contextual Image Generation

- Given a Visual Story, generate a relevant image
- Need to condition on long, temporally dependent text
- (Optionally) Condition on image inputs interleaved within the text

Visual Storytelling (Huang et al., 2016)

Evaluation: Contextual Image Generation

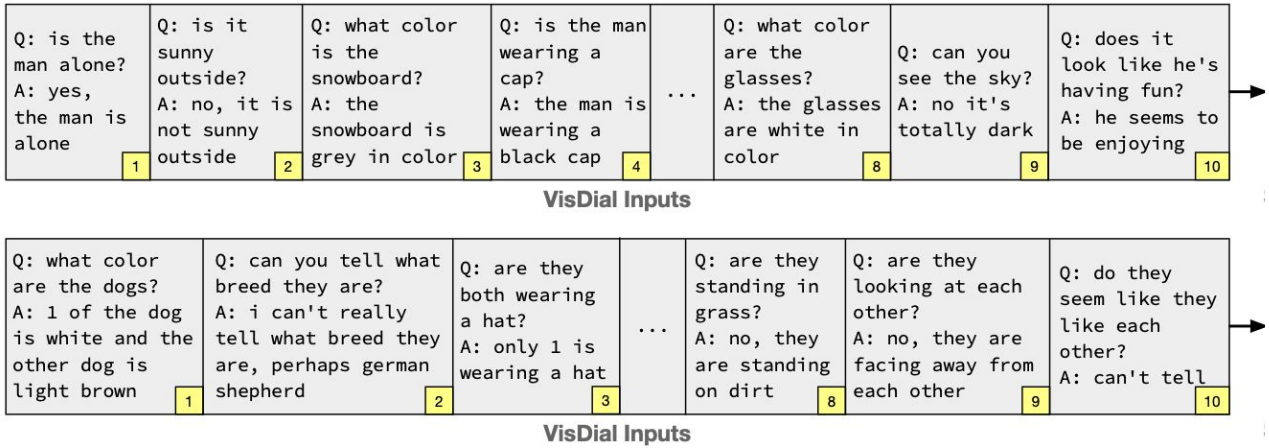
Model	CLIP Similarity (↑)			LPIPS (↓)		
	1 caption	5 captions	5 caps, 4 images	1 caption	5 captions	5 caps, 4 images
GLIDE [34]	0.582	0.591	-	0.753	0.745	-
Stable Diffusion [43]	**0.592** ±0.0007	0.598 ± 0.0006	-	0.703 ±0.0003	0.704 ± 0.0004	-
GILL	0.581 ±0.0005	**0.612** ±0.0011	**0.641** ±0.0011	**0.702** ±0.0004	**0.696** ±0.0008	**0.693** ±0.0008

- Our model outperforms Stable Diffusion on longer input contexts
- This is despite GILL (essentially) distilling from SD!
- GILL benefits from the abilities of the LLM (sensitivity to longer inputs, word orderings, in-context learning)

Visual Storytelling (Huang et al., 2016)

Evaluation: Contextual Image Generation

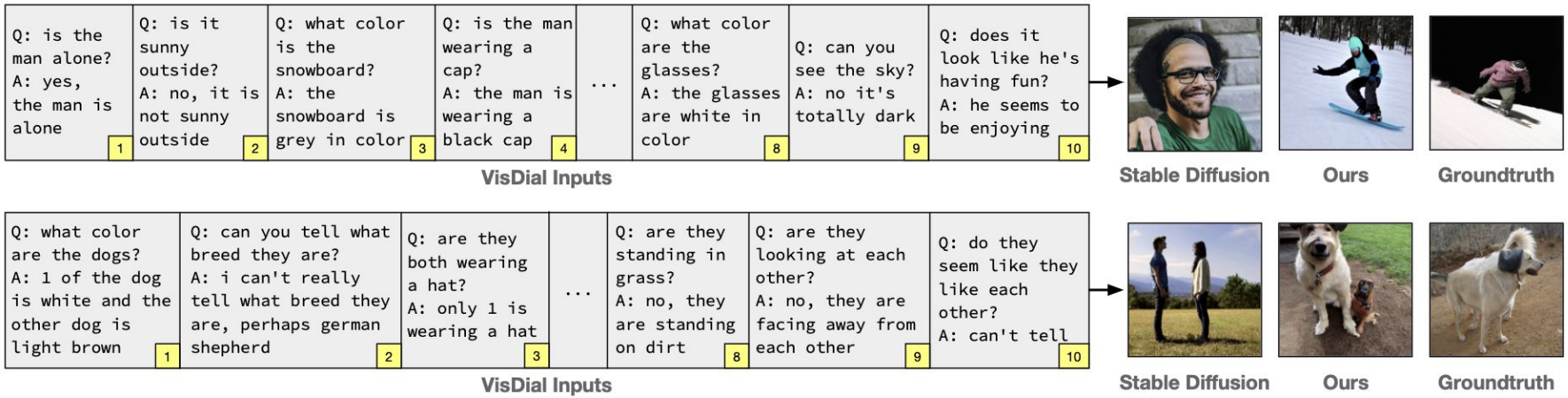
- Given a Visual Dialogue, generate a relevant image



Visual Dialog (Das et al., 2017)

Evaluation: Contextual Image Generation

- Given a Visual Dialogue, generate a relevant image
- Need to condition on long dialogue-like text (OOD with finetuning data)



Visual Dialog (Das et al., 2017)

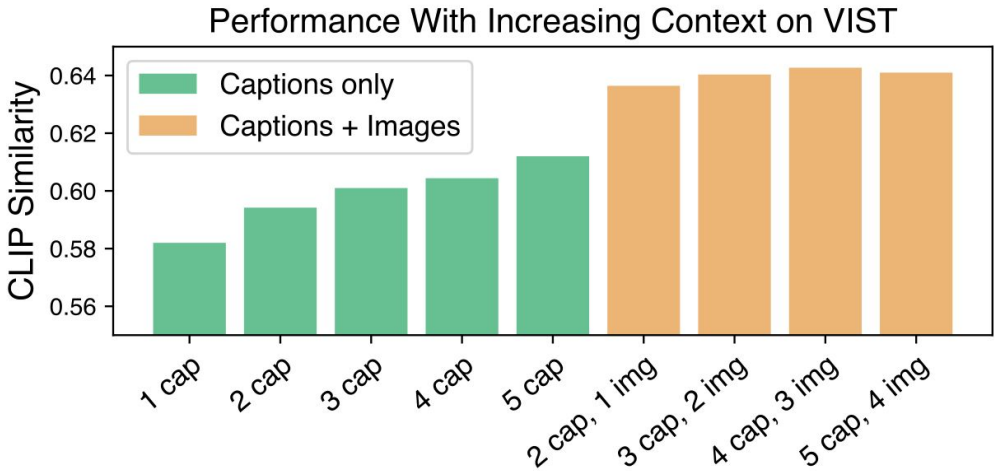
Evaluation: Contextual Image Generation

	CLIP Similarity (↑)			LPIPS (↓)		
Model	1 round	5 rounds	10 rounds	1 round	5 rounds	10 rounds
GLIDE [34]	**0.562**	0.595	0.587	0.800	0.794	0.799
Stable Diffusion [43]	0.552 ±0.0015	**0.629** ±0.0015	0.622 ±0.0012	**0.742** ±0.0010	0.722 ±0.0012	0.723 ±0.0008
GILL	0.528 ±0.0014	0.621 ±0.0009	**0.645** ±0.0010	**0.742** ±0.0022	**0.718** ±0.0028	**0.714** ±0.0006

Visual Storytelling (Huang et al., 2016)

The Effect of Context

Multi-modal context is **worth more** than uni-modal context, producing more relevant generation results.



GILLMapper: A Stronger LLM-to-Generator Mapping

Image generators require **denser** input sequences. Linear mappings are insufficient.

Model	CC3M FID (↓)	VIST CLIP Sim (↑)
Stable Diffusion [43]	**13.94**	0.598
Ours + Linear	15.50	0.500
Ours + 3-layer MLP	15.33	0.502
Ours + Transformer Encoder	16.30	0.605
Ours + GILLMapper	15.31	**0.641**

Other Abilities: Text-to-Image Generation

Stable Diffusion

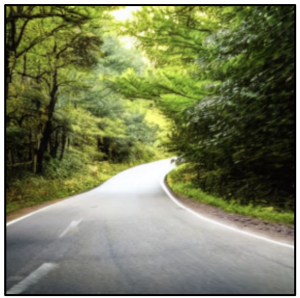
Ours

“A dignified beaver wearing glasses, a vest, and colorful neck tie. He stands next to a tall stack of books in a library.”

Stable Diffusion

Ours

“Snow mountain and tree reflection in the lake”



Stable Diffusion

Ours

“A drop-top sports car coming around a bend in the road”

Stable Diffusion

Ours

“a group of penguins in a snowstorm”

Other Abilities: Image Refinement

I want to paint a landscape with mountains and a river. Can you help me find some inspiration?

I'm not sure

Can you show me some with horses?

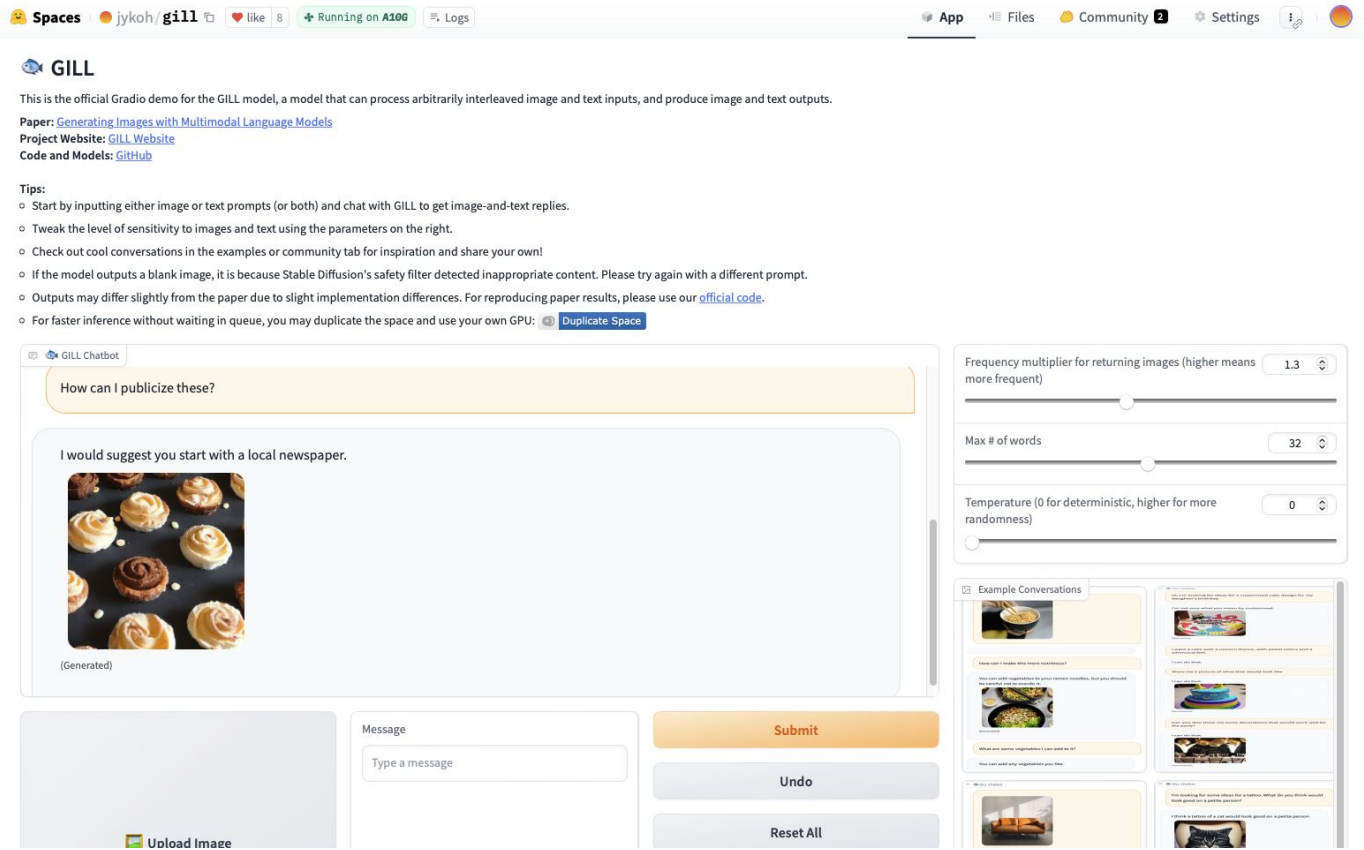
What about something more abstract?

Future Work

- **Train on more diverse data**
  - CC3M is small by modern standards – we would get a lot more from training on LAION
  - Training on interleaved image-text data would also likely help a lot
  - GILLMapper will likely be more aligned to SD
- **Apply to even larger LLMs and stronger visual models**
  - We use a 7B LLM, but you can likely train a 13B LLM with a few A6000 GPUs
- **Use a finetuned LLM**
  - For example, instruction finetuned, or dialogue finetuned
- **Perform more explicit image conditioning**
  - May allow the model to be better at tasks such as image editing

Try the model!

huggingface.co/spaces/jykoh/gill



Thanks!

jykoh@cmu.edu
jykoh.com/gill