

Generating Images with Multimodal Language Models

Jing Yu Koh, Daniel Fried, Ruslan Salakhutdinov

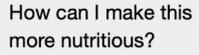
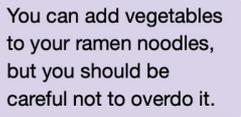
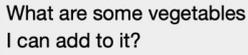
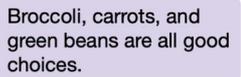
NeurIPS 2023

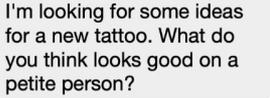
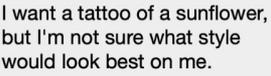
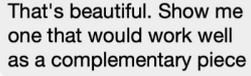
Carnegie
Mellon
University

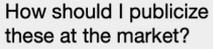
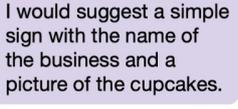


Can we *ground* text-only LLMs to pretrained visual encoders and decoders?

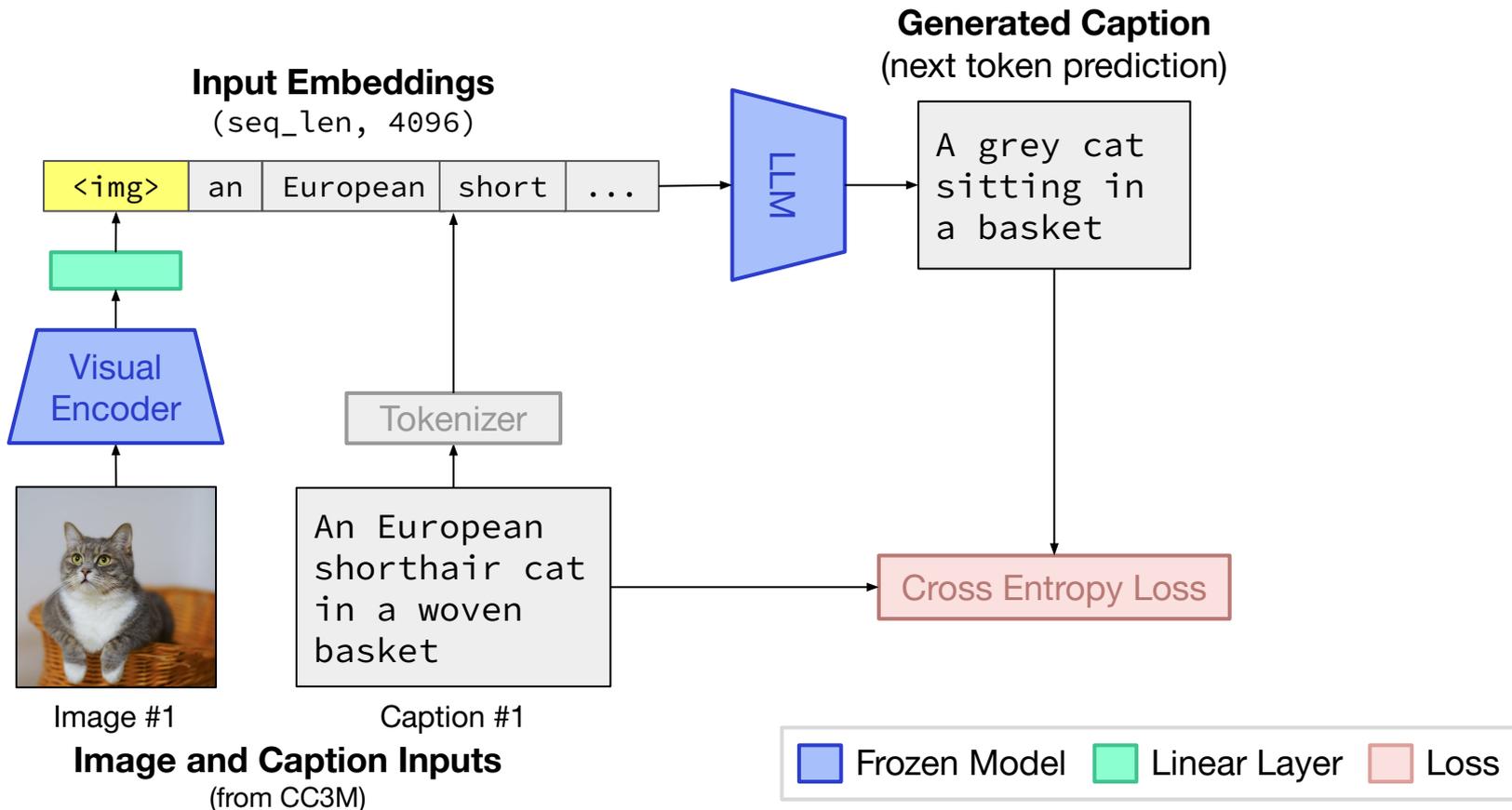
Generating Images with Large Language Models

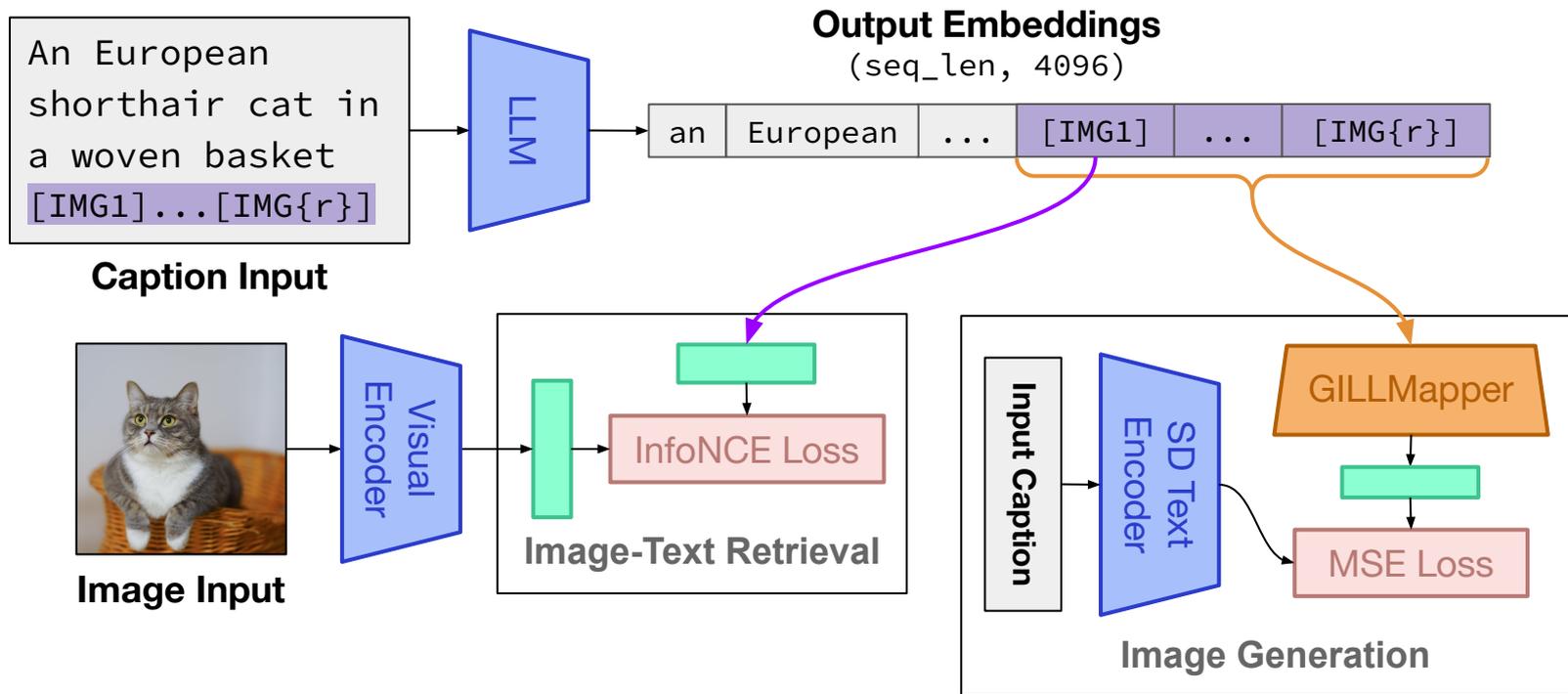
     

Learning to *Process* Images

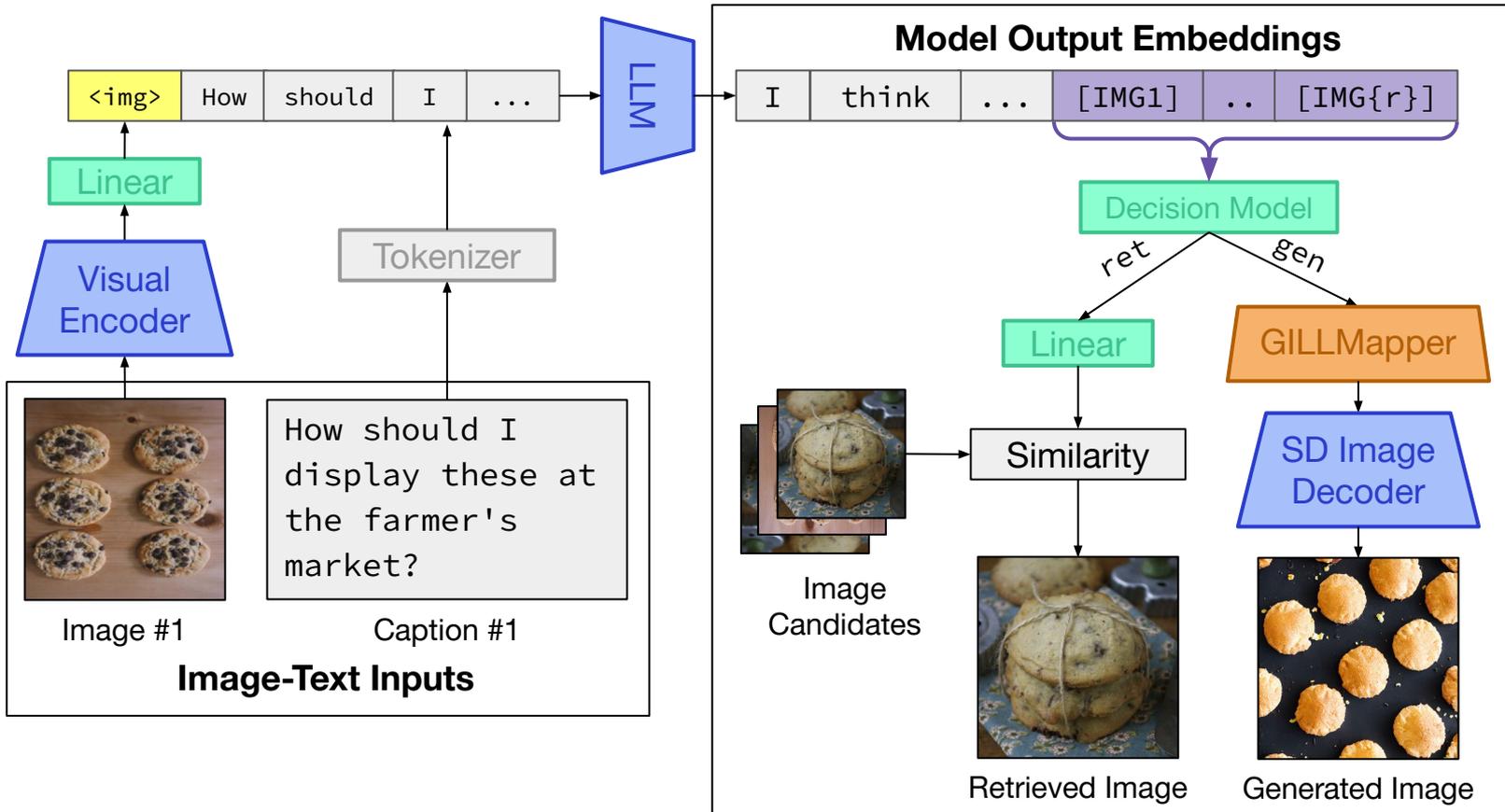


Learning to *Produce* Images



Legend:

- Frozen Model (Blue trapezoid)
- GILLMapper (Orange trapezoid)
- Linear Layer (Green rectangle)
- Loss (Red rectangle)



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
- Need to condition on long, temporally dependent text
- (Optionally) Condition on image inputs interleaved within the text

Once while I was on vacation in this nice brick hotel		I woke up and took my dog Trixie for a walk.		Trixie ran around and enjoyed the fresh air.		We had lots of fun playing fetch together		After a while she got tired and had to take a rest.				
Image and Text Inputs									→	Stable Diffusion	Ours	Groundtruth

Evaluation: Contextual Image Generation

Model	CLIP Similarity (\uparrow)			LPIPS (\downarrow)		
	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)

Evaluation: Contextual Image Generation

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

Q: is the man alone? A: yes, the man is alone 1	Q: is it sunny outside? A: no, it is not sunny outside 2	Q: what color is the snowboard? A: the snowboard is grey in color 3	Q: is the man wearing a cap? A: the man is wearing a black cap 4	...	Q: what color are the glasses? A: the glasses are white in color 8	Q: can you see the sky? A: no it's totally dark 9	Q: does it look like he's having fun? A: he seems to be enjoying 10
-------------------------------------------------------	----------------------------------------------------------------	---------------------------------------------------------------------------	------------------------------------------------------------------------	-----	--------------------------------------------------------------------------	---------------------------------------------------------	---------------------------------------------------------------------------

VisDial Inputs



Stable Diffusion



Ours



Groundtruth

Q: what color are the dogs? A: 1 of the dog is white and the other dog is light brown 1	Q: can you tell what breed they are? A: i can't really tell what breed they are, perhaps german shepherd 2	Q: are they both wearing a hat? A: only 1 is wearing a hat 3	...	Q: are they standing in grass? A: no, they are standing on dirt 8	Q: are they looking at each other? A: no, they are facing away from each other 9	Q: do they seem like they like each other? A: can't tell 10
-----------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------	-----	-------------------------------------------------------------------------	----------------------------------------------------------------------------------------	-------------------------------------------------------------------

VisDial Inputs



Stable Diffusion



Ours

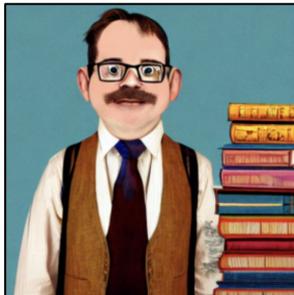


Groundtruth

Evaluation: Contextual Image Generation

Model	CLIP Similarity (\uparrow)			LPIPS (\downarrow)		
	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 \pm 0.0015	0.629 \pm 0.0015	0.622 \pm 0.0012	0.742 \pm 0.0010	0.722 \pm 0.0012	0.723 \pm 0.0008
GILL	0.528 \pm 0.0014	0.621 \pm 0.0009	0.645 \pm 0.0010	0.742 \pm 0.0022	0.718 \pm 0.0028	0.714 \pm 0.0006

Other Abilities: Text-to-Image Generation



Stable Diffusion

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



Ours

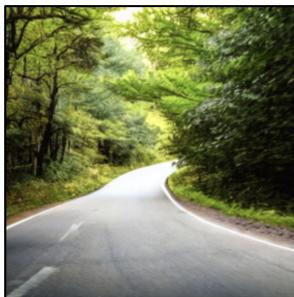


Stable Diffusion

“Snow mountain and tree reflection in the lake”



Ours



Stable Diffusion

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



Ours



Stable Diffusion

“a group of penguins in a snowstorm”



Ours

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?



Thanks!

jykoh@cmu.edu

jykoh.com/gill