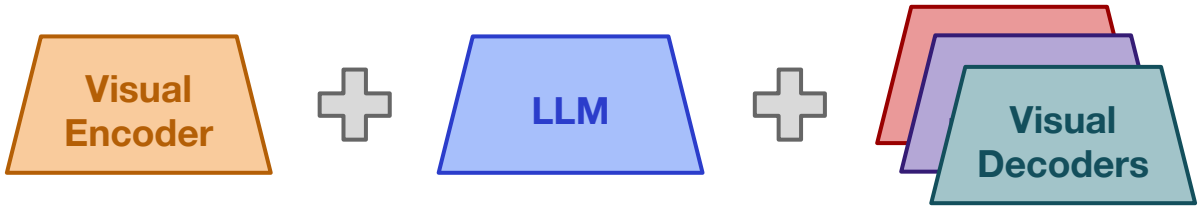
Generating Images with Multimodal Language Models

Jing Yu Koh, Daniel Fried, Ruslan Salakhutdinov

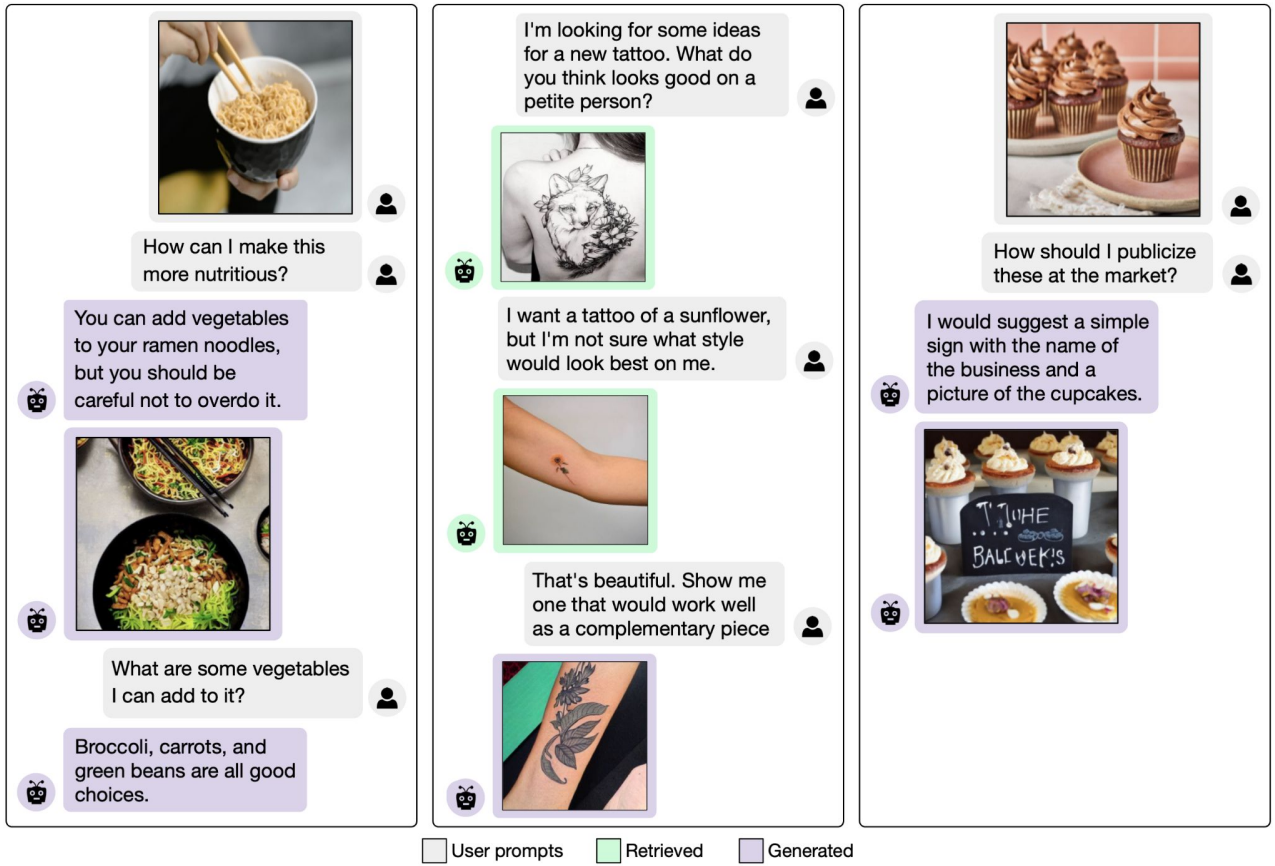
NeurIPS 2023



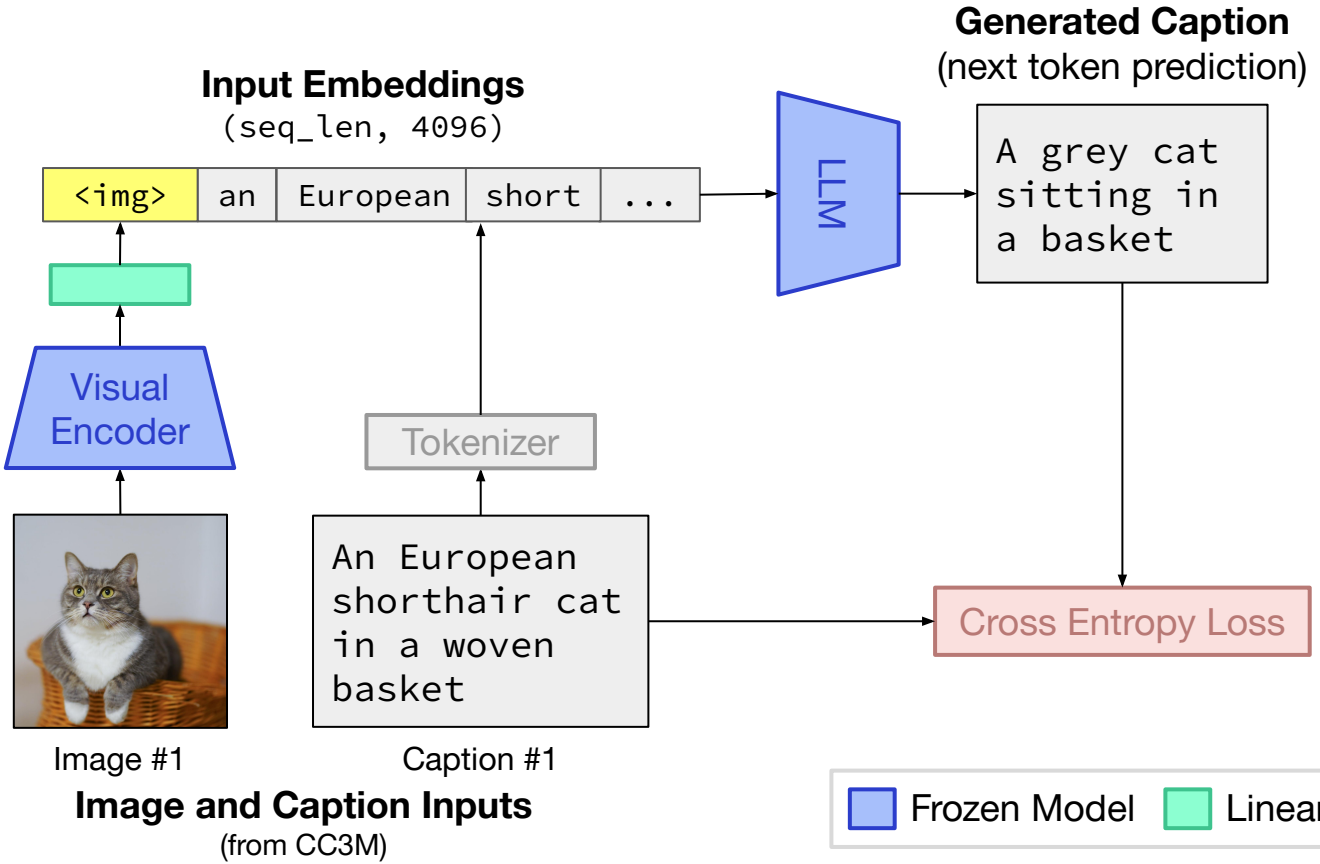
Can we *ground* text-only LLMs to pretrained visual encoders and decoders?

GILL

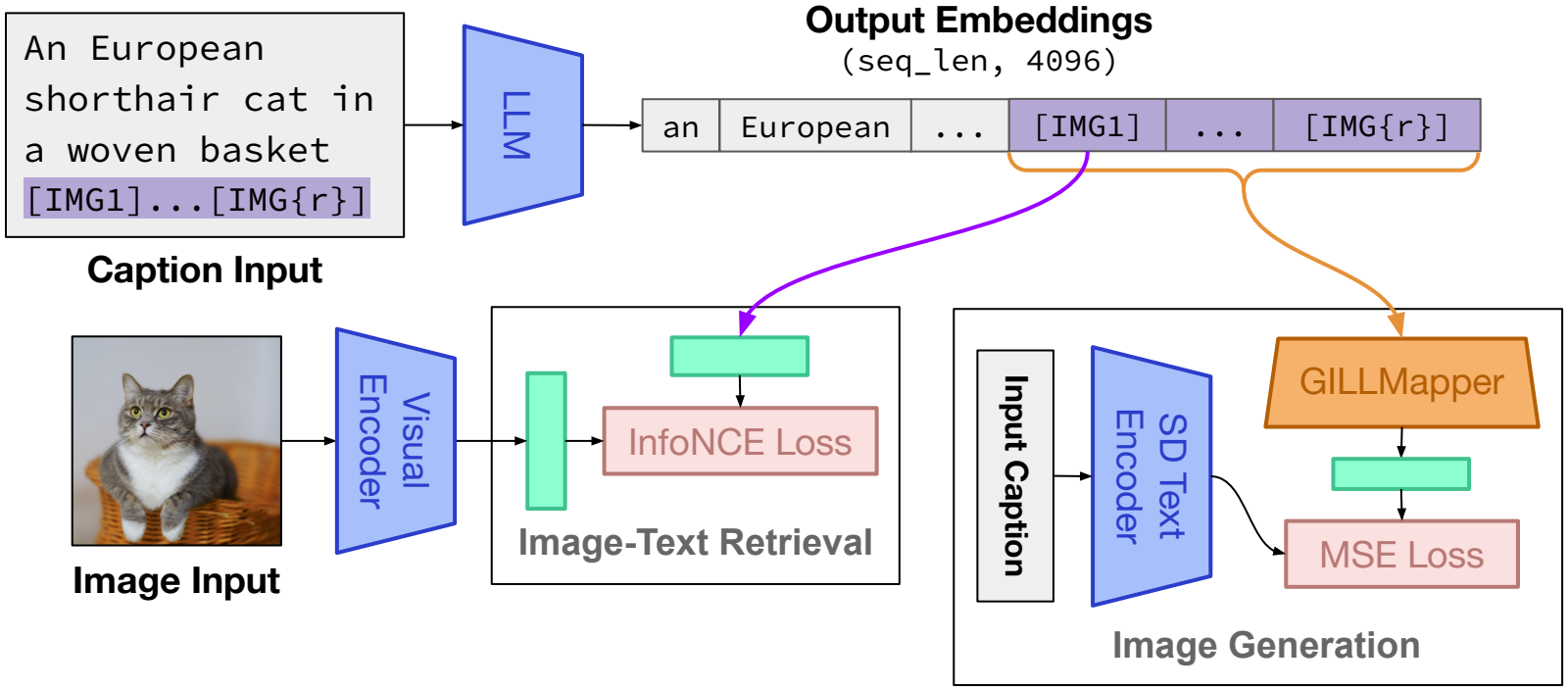
Generating Images with Large Language Models

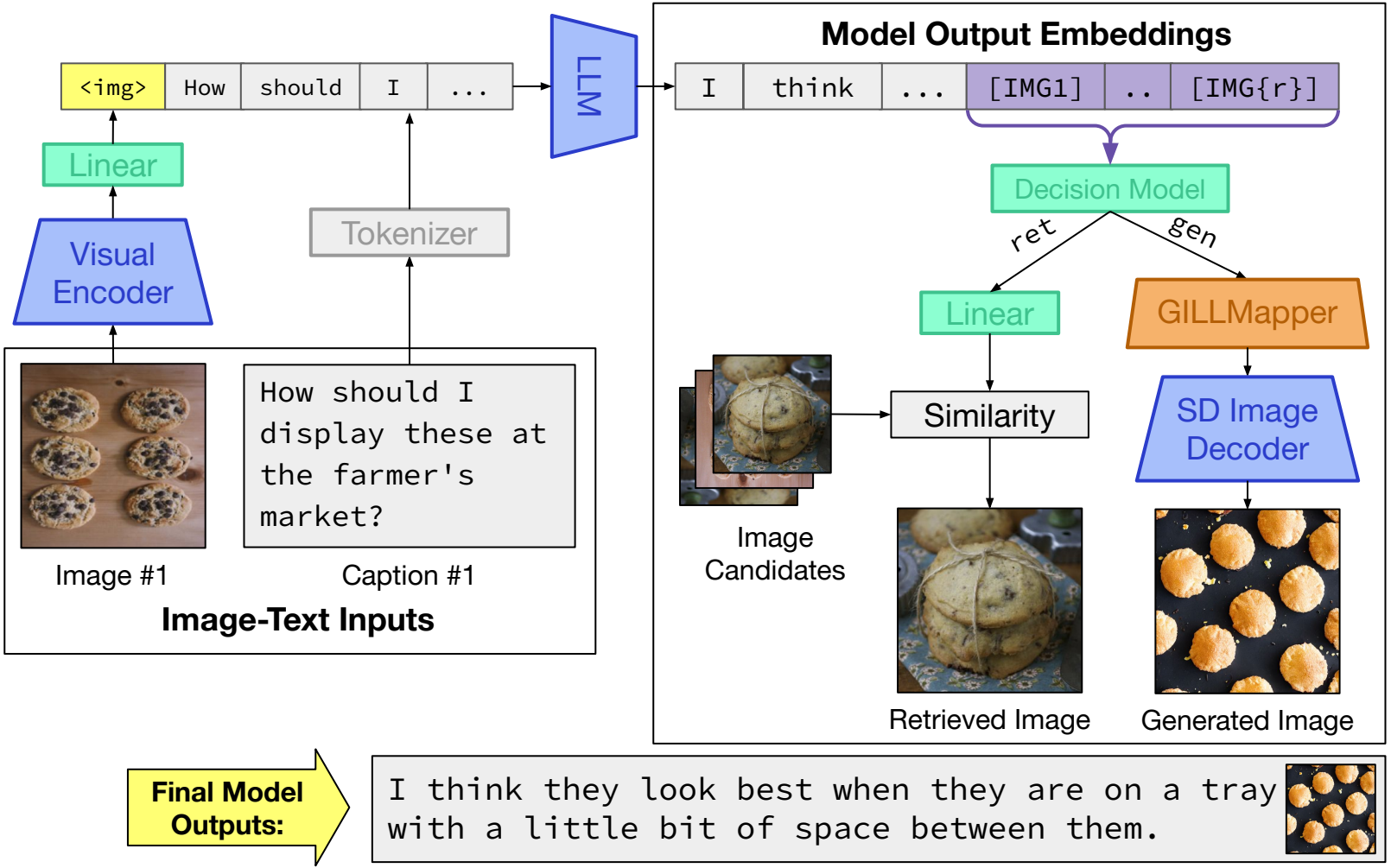


Learning to *Process* Images



Learning to *Produce* Images




<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
- 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

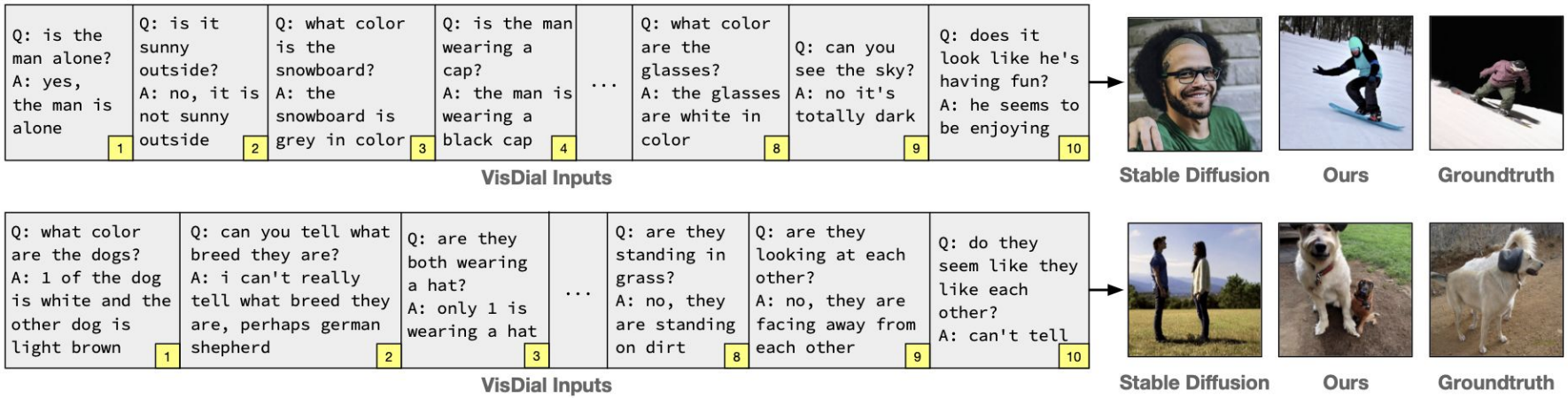
Model	CLIP Similarity (↑)			LPIPS (↓)		
	1 caption	5 captions	5 caps, 4 images	1 caption	5 captions	5 caps, 4 images
GLIDE [34]	0.582	0.591	-	0.753	0.745	-
Stable Diffusion [43]	**0.592** ±0.0007	0.598 ± 0.0006	-	0.703 ±0.0003	0.704 ± 0.0004	-
GILL	0.581 ±0.0005	**0.612** ±0.0011	**0.641** ±0.0011	**0.702** ±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
- Need to condition on long dialogue-like text (OOD with finetuning data)



Visual Dialog (Das et al., 2017)

Evaluation: Contextual Image Generation

Model	CLIP Similarity (↑)			LPIPS (↓)		
	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)

Other Abilities: Text-to-Image Generation

Stable Diffusion

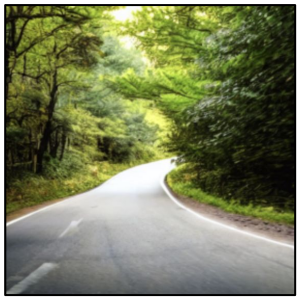
Ours

“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?

Thanks!

jykoh@cmu.edu
jykoh.com/gill