Generating Contextual Images for Long-Form Text

Avijit Mitra[1], Nalin Gupta[2], Chetan Nagaraj Naik[2], Abhinav Sethy[2], Kinsey Bice[2], Zeynab Raeesy[2]

[1]University of Massachusetts Amherst [2]Amazon

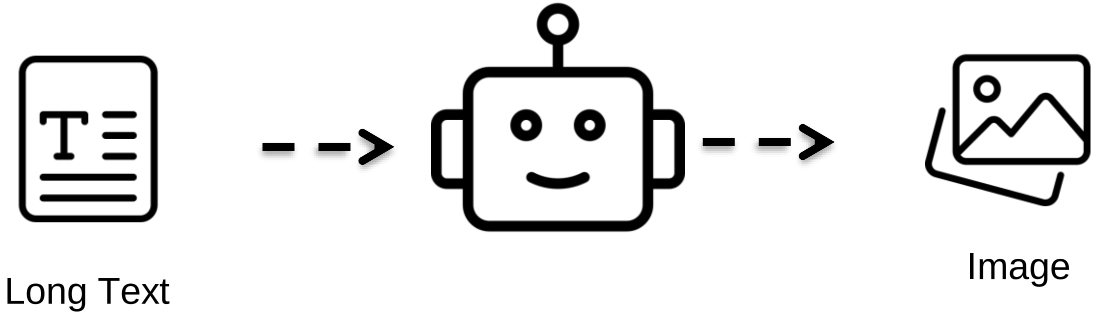
Presentation outline

- Problem Statement & Motivation
- Literature Review
- Dataset
- Experiments
- Metrics
- Results
- Qualitative Examples
- Summary

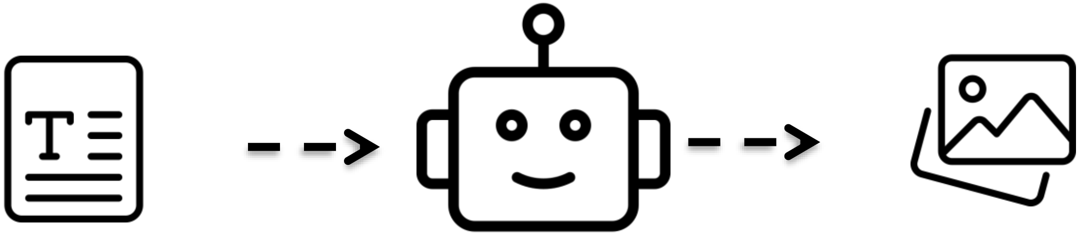
Presentation outline

- Problem Statement & Motivation
- Literature Review
- Dataset
- Experiments
- Metrics
- Results
- Qualitative Examples
- Summary

Contextual Images from Long-form Text



Contextual Images from Long-form Text

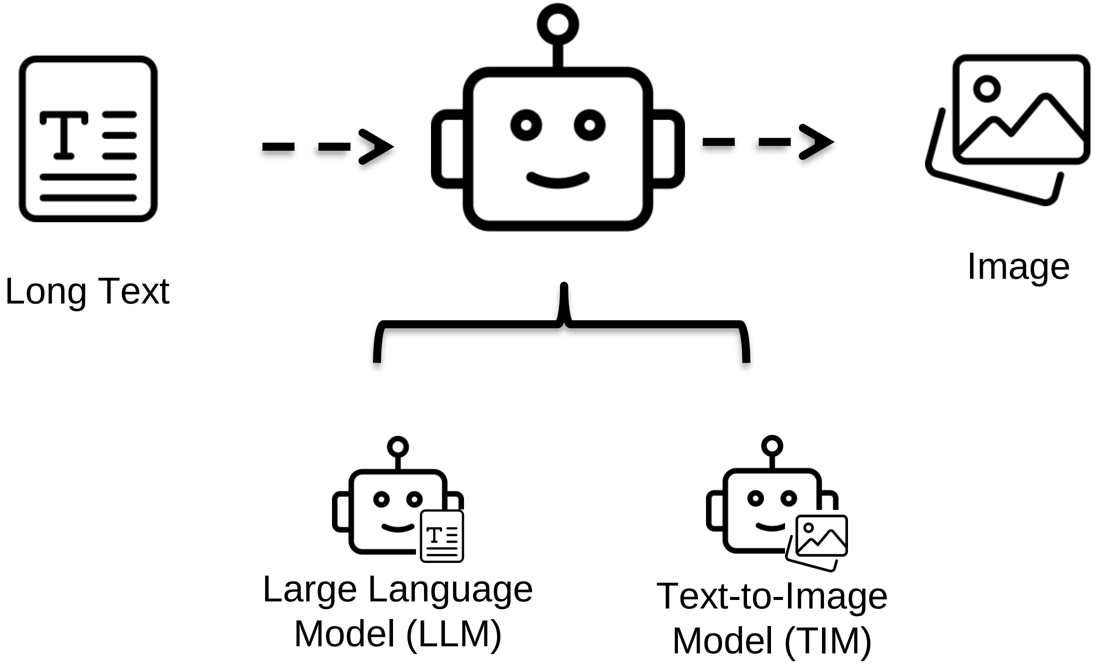


Long Text

Image

The Department of Justice and Public Safety in the Canadian province of New Brunswick was formed when Premier Brian Gallant restructured government departments in 2016. …. Largely created from the former Department of the Solicitor General, …. The department is headed by a Minister of Justice and Public Safety who also continues to hold the title of Solicitor General of New Brunswick (French: *Ministre de la Sécurité Publique et Solliciteur Général*).

Contextual Images from Long-form Text



LLM vs TIM

LLM

- ✓ Can process large chunk of text
- ✓ Can generate coherent text given suitable prompt (and sometimes a few examples)
- ✗ Can not generate image

TIM

- ✓ Can generate image conditioned on concise text prompt
- ✗ Can not process large chunk of text
- ✗ Can not generate text
- ✗ Lack the strong reasoning capability like LLMs

Motivations

- Scopes of research:
  - Lack of open-sourced reliable systems for image generation from long-form text.
  - Limited work on how to evaluate such systems.
- Applications:
  - General: Wikipedia contribution, story writers
  - Industrial: Can be modified to fit various use cases on smart displays

Presentation outline

- Problem Statement & Motivation
- Literature Review
- Dataset
- Experiments
- Metrics
- Results
- Qualitative Examples
- Summary

Related Works

- Synthesizing multimodal content.

Model	Model Configuration						Image-Text Data		Visual Instruction Data	
	VE	LLM	Adapt	ToP	TuP	# Token	Source	Size	Source	Size
BLIP2	ViT-g/14†	FlanT5-XL†	Q-Former	4B	107M	32	CC*-VG-SBU-L400	129M	-	-
LLaVA	ViT-L/14†	Vicuna	FC layer	7B	7B	256	CC3M	595K	LLaVA-I	158K
LA-V2	ViT-L/14†	LLaMA†	B-Tuning	7B	63.1M	10	L400	200M	LLaVA-I	158K
MiniGPT-4	BLIP2-VE†	Vicuna†	FC layer	7B	3.1M	32	CC-SBU-L400	5M	CC+ChatGPT	3.5K
mPLUG-Owl	ViT-L/14	LLaMA†	LoRA	7B	388M	65	CC*-CY-L400	204M	LLaVA-I	158K
Otter	ViT-L/14†	LLaMA†	Resampler	9B	1.3B	64	-	-	LLaVA-I	158K
InstructBLIP	ViT-g/14†	Vicuna†	Q-Former	7B	107M	32	-	-	QA*	16M
VPGTrans	ViT-g/14†	Vicuna†	Q-Former	7B	107M	32	COCO-VG-SBU-LC	13.8M	CC+ChatGPT	3.5K

*Table from [1]

More recent models: ImageBind, Multimodal-GPT, KOSMOS-2 etc.

[1] Xu, Peng, et al. "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models." *arXiv preprint arXiv:2306.09265* (2023).

Related Works

- Generating Images from Text
  - GANs
    - Conditional GAN, multi-stage GAN, attention GAN, cross-modal contrastive GAN, VQGAN etc.
  - Transformer-based decoders
    - Cogview, Maskgit etc.
  - Diffusion models
    - Models from Stable diffusion family, GLIDE etc.

Related Works

- Generating Images from Long-form Text
  - GILL [2]
    - Integrates LLM with TIM
    - Do not focus on longer text input.

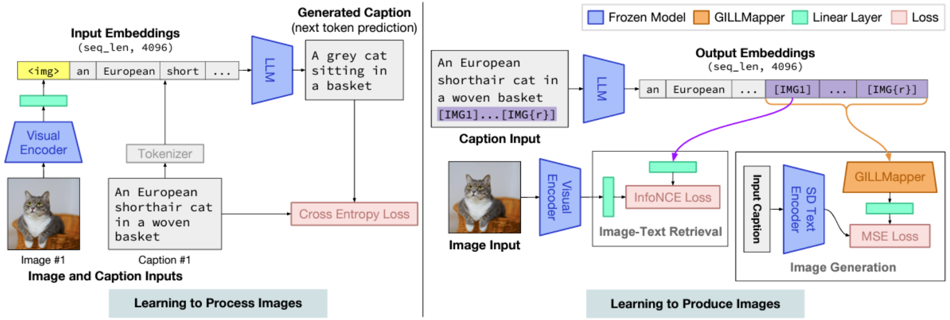


Figure : GILL model architecture overview. It is trained with a captioning loss to learn to process images (left), and losses for image retrieval and image generation to learn to produce images (right).

[2] Koh, Jing Yu, Daniel Fried, and Ruslan Salakhutdinov. "Generating images with multimodal language models." *arXiv preprint arXiv:2305.17216* (2023).

Related Works

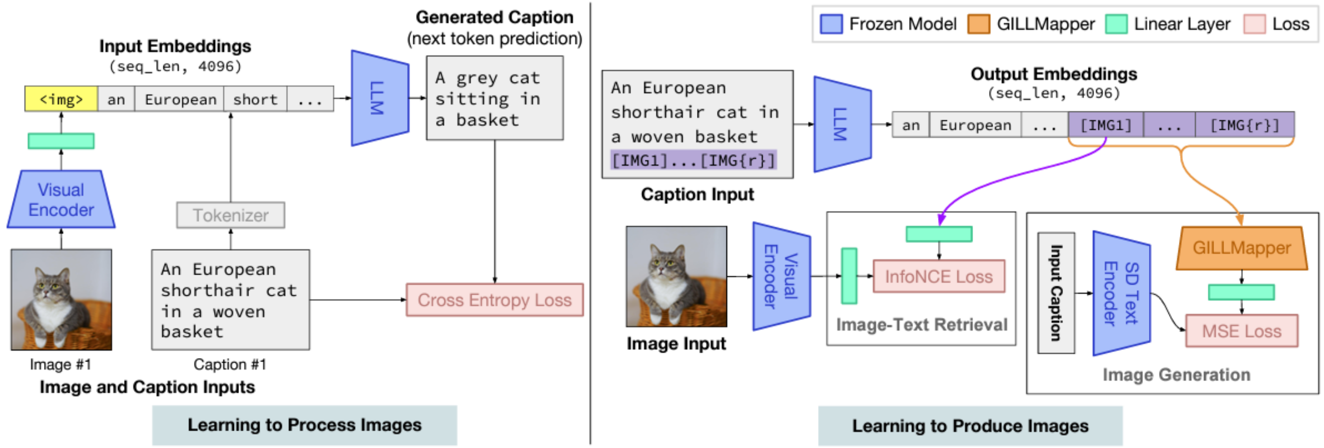


Figure : GILL model architecture overview. It is trained with a captioning loss to learn to process images (left), and losses for image retrieval and image generation to learn to produce images (right).

- Model and codes not publicly available

LLM: OPT-6.7B
Visual Encoder: CLIP ViT-L
TIM: SD v1.5

[2] Koh, Jing Yu, Daniel Fried, and Ruslan Salakhutdinov. "Generating images with multimodal language models." *arXiv preprint arXiv:2305.17216* (2023).
[3] Aghajanyan, Armen, et al. "Cm3: A causal masked multimodal model of the internet." *arXiv preprint arXiv:2201.07520* (2022).

Related Works

- Generating Images from Long-form Text
  - GILL [2]
    - Integrates LLM with TIM
    - Do not focus on longer text input.
  - CM3 [3]
    - Requires restructuring all tasks in HTML format.
    - Model and codes not publicly available

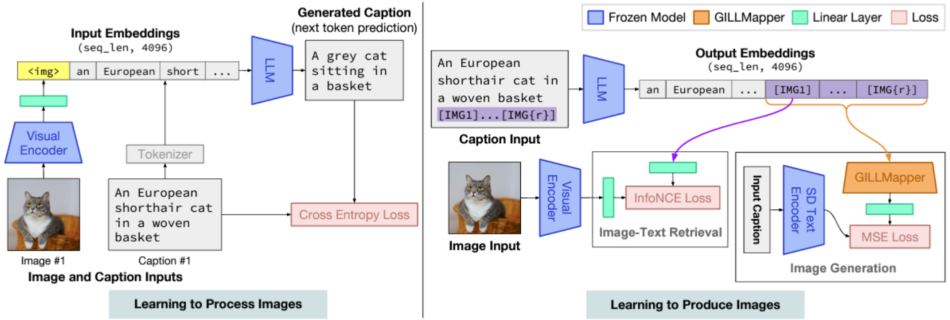


Figure : GILL model architecture overview. It is trained with a captioning loss to learn to process images (left), and losses for image retrieval and image generation to learn to produce images (right).

[2] Koh, Jing Yu, Daniel Fried, and Ruslan Salakhutdinov. "Generating images with multimodal language models." *arXiv preprint arXiv:2305.17216* (2023).
[3] Aghajanyan, Armen, et al. "Cm3: A causal masked multimodal model of the internet." *arXiv preprint arXiv:2201.07520* (2022).

Presentation outline

- Problem Statement & Motivation
- Literature Review
- Dataset
- Experiments
- Metrics
- Results
- Qualitative Examples
- Summary

Dataset

Dataset	Description	#English Instances
WIT	Wikipedia sections along with images, precursor of WikiWeb2M	~5.4M
WikiWeb2M	A superset of WIT with more content	~11.7M
MMC4	Common Crawl text data with interleaved images	~101.2M
CC3M	A collection of image-caption pairs	~3.37M
MS-COCO	Human annotated image caption pairs	~330k
VisDial	Visual dialogue data, contains an image with 10 Q/A pairs	~133k
LAION-400M	CLIP-filtered image-text pairs, with CLIP embeddings	~400M
LAION-5B	A superset of LAION-400M	~5.85B

Dataset

Dataset	Description	#English Instances
WIT [4]	Wikipedia sections along with images, precursor of WikiWeb2M	~5.4M
WikiWeb2M	A superset of WIT with more content	~11.7M
MMC4	Common Crawl text data with interleaved images	~101.2M
CC3M	A collection of image-caption pairs	~3.37M
MS-COCO	Human annotated image caption pairs	~330k
VisDial	Visual dialogue data, contains an image with 10 Q/A pairs	~133k
LAION-400M	CLIP-filtered image-text pairs, with CLIP embeddings	~400M
LAION-5B	A superset of LAION-400M	~5.85B

[4] Srinivasan, Krishna, et al. "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning." *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021.

Dataset

Dataset	Description	#English Instances
WIT [4]	Wikipedia sections along with images, precursor of WikiWeb2M	~5.4M
	ntent	~11.7M
	terleaved ima	
	irs	
	n pairs	
	n image with	
	ith CLIP emb	
		~5.85B

Includes:
- Page Title
- Section Title
- Image Caption
- Page Description
- Corresponding Image
- Page URL
- Image URL etc.

[7] Srinivasan, Krishna, et al. "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning." *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021.

Dataset

- Challenges:

    1. Images are often supplementary to the page descriptions.

Dataset

- Challenges:

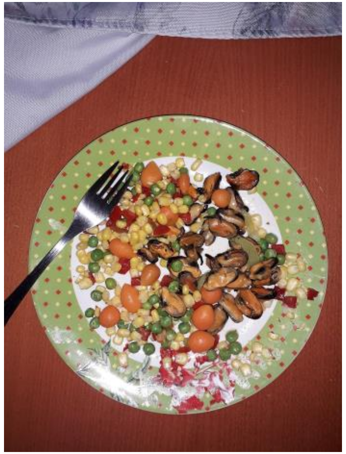
  1. Images are often supplementary to the page descriptions.

Barbastathis (Greek: Μπαρμπαστάθης) is the name of a Greek brand of frozen vegetables, owned today by CVC Capital Partners.

It was founded in 1969 in Thessaloniki by Giannis Michailidis from Drama.

In 1991 it entered the Athens Stock Exchange and in 1994 it was by bought by Delta dairy company (Daskalopoulos).[1]

As of 2017 it is the leading brand of frozen vegetables, found mostly in Greek supermarkets. The company maintains a factory in the industrial area of Sindos, Thessaloniki.[2]



Mussels dish with *Barbastathis* corn salad

Dataset

- Challenges:

  1. Images are often supplementary to the page descriptions.

  2. Not all pages are good candidates for multimodal generation.

Dataset

- Challenges:

    1. Images are often supplementary to the page descriptions.

    2. Not all pages are good candidates for multimodal generation.

Jerry Gana, is a Nigerian scholar, politician and one time senator of the Federal Republic of Nigeria in 1983 then Director for the Directorate of Food, Roads and Infrastructure (DFRRI).He was the director of the Mass Mobilization for Social Justice and Economic Recovery, popularly known as MAMSER under Ibrahim Babangida,[1] then Minister of Agriculture and Natural Resources, in the Interim National Government under Ernest Shonekan.[1] Later he became Minister of Information and Culture under General Sani Abacha, then Minister of Corporation and Integration in Africa under Olusegun Obasanjo as well as being Minister of Information and national Orientation. He also served as Political Adviser to Olusegun Obasanjo, before announcing plans to run for president in June 2006.[2]

Professor Gana signature

Dataset

- Challenges:

    1. Images are often supplementary to the page descriptions.

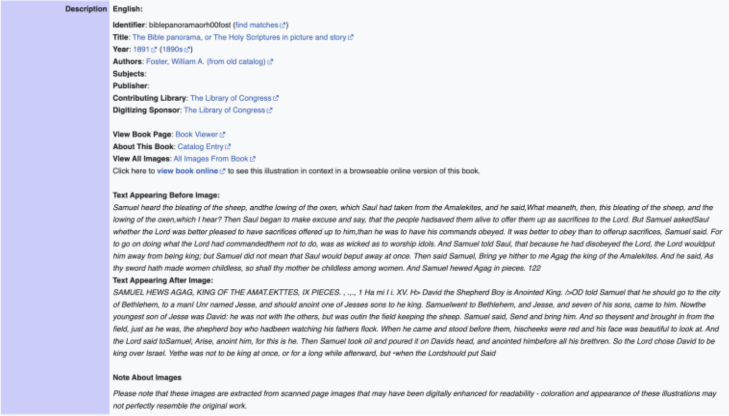
    2. Not all pages are good candidates for multimodal generation.

    3. Lacks good image captions, often noisy or uninformative.

Dataset

- Challenges:

  1. Images are often supplementary to the page descriptions.
  2. Not all pages are good candidates for multimodal generation.
  3. Lacks good image captions, often noisy or uninformative.



Description **English:**

Identifier: biblepanoramaorh00fost (find matches)
Title: The Bible panorama, or The Holy Scriptures in picture and story
Year: 1891 (1890s)
Authors: Foster, William A. (from old catalog)
Subjects:
Publisher:
Contributing Library: The Library of Congress
Digitizing Sponsor: The Library of Congress

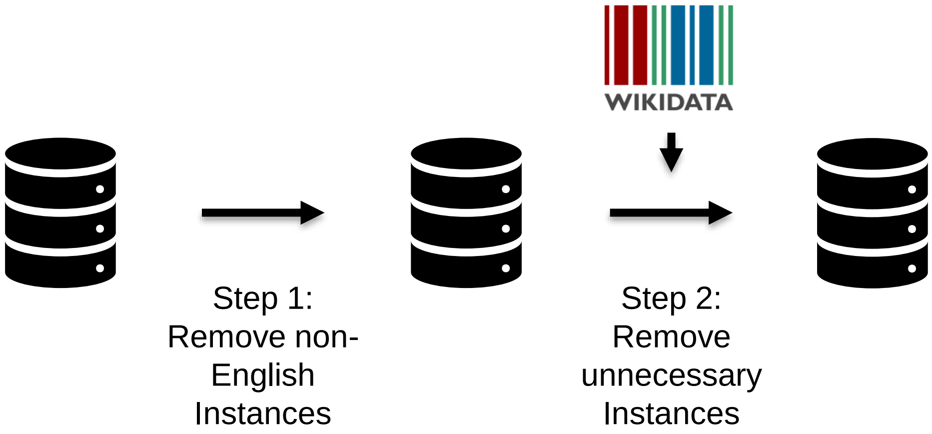
View Book Page: Book Viewer
About This Book: Catalog Entry
View All Images: All Images From Book
Click here to **view book online** to see this illustration in context in a browseable online version of this book.

Text Appearing Before Image:
Samuel heard the bleating of the sheep, andthe lowing of the oxen, which Saul had taken from the Amalekites, and he said,What meaneth, then, this bleating of the sheep, and the lowing of the oxen,which I hear? Then Saul began to make excuse and say, that the people hadsaved them alive to offer them up as sacrifices to the Lord. But Samuel askedSaul whether the Lord was better pleased to have sacrifices offered up to him,than he was to have his commands obeyed. It was better to obey than to offerup sacrifices, Samuel said. For to go on doing what the Lord had commandedthem not to do, was as wicked as to worship idols. And Samuel told Saul, that because he had disobeyed the Lord, the Lord wouldput him away from being king; but Samuel did not mean that Saul would beput away at once. Then said Samuel, Bring ye hither to me Agag the king of the Amalekites. And he said, As thy sword hath made women childless, so shall thy mother be childless among women. And Samuel hewed Agag in pieces. 122
Text Appearing After Image:
SAMUEL HEWS AGAG, KING OF THE AMAT.EKTTES, IX PIECES. , .,., 1 Ha mi I i. XV. H> David the Shepherd Boy is Anointed King. />OD told Samuel that he should go to the city of Bethlehem, to a manI Unr named Jesse, and should anoint one of Jesses sons to he king. Samuelwent to Bethlehem, and Jesse, and seven of his sons, came to him. Nowthe youngest son of Jesse was David: he was not with the others, but was outin the field keeping the sheep. Samuel said, Send and bring him. And so theysent and brought in from the field, just as he was, the shepherd boy who hadbeen watching his fathers flock. When he came and stood before them, hischeeks were red and his face was beautiful to look at. And the Lord said toSamuel, Arise, anoint him, for this is he. Then Samuel took oil and poured it on Davids head, and anointed himbefore all his brethren. So the Lord chose David to be king over Israel. Yethe was not to be king at once, or for a long while afterward, but •when the Lordshould put Said

Note About Images

Please note that these images are extracted from scanned page images that may have been digitally enhanced for readability - coloration and appearance of these illustrations may not perfectly resemble the original work.

Dataset



Datasets	#Train	#Dev	#Test
WIT	37,046,386	261,024	210,166
WIT (after step 1)	5,407,014	45,405	33,070
WIT (after step 2)	359,822*	31,381	22,554

*10% of available data

Presentation outline

- Problem Statement & Motivation
- Literature Review
- Dataset
- Experiments
- Metrics
- Results
- Qualitative Examples
- Summary

Experiments

2 Groups:

1. Zero-shot
    1. SD_{OPT} (OPT+SD)
    2. SD_{Vicuna} (Vicuna + SD)
    3. GILL
2. Fine-tuning
    1. GILL
    2. GILL with Vicuna

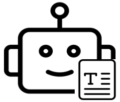
Experiments

2 Groups:

1. Zero-shot
   1. SD_{OPT} (OPT+SD)
   2. $\text{SD}_{\text{Vicuna}}$ (Vicuna + SD)
   3. GILL
2. Fine-tuning
   1. GILL
   2. GILL with Vicuna

Summarize into one sentence that can be used as the caption of a corresponding image: The Department of Justice and Public Safety in the Canadian province of New Brunswick was formed when Premier Brian Gallant restructured government departments in 2016. …. Largely created from the former Department of the Solicitor General, …. The department is headed by a Minister of Justice and Public Safety who also continues to hold the title of Solicitor General of New Brunswick (French: *Ministre de la Sécurité Publique et Solliciteur Général*).
Answer:

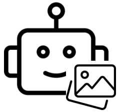
Input Long Text with prompt



Large Language Model (LLM)

2016 restructuring of the Department of Justice and Public Safety in New Brunswick created … to the Department of Finance

Summary



Text-to-Image Model (TIM)

Image Output

Color code:
<Instruction> <Input> <Trigger_prompt>

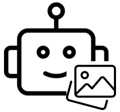
Experiments

2 Groups:

1. Zero-shot
    1. SD_{OPT} (OPT+SD)
    2. SD_{Vicuna} (Vicuna + SD)
    3. GILL
2. Fine-tuning
    1. GILL
    2. GILL with Vicuna

The Department of Justice and Public Safety in the Canadian province of New Brunswick was formed when Premier Brian Gallant restructured government departments in 2016. …. Largely created from the former Department of the Solicitor General, …. The department is headed by a Minister of Justice and Public Safety who also continues to hold the title of Solicitor General of New Brunswick (French: *Ministre de la Sécurité Publique et Solliciteur Général*).

Input Long Text with prompt



Pretrained GILL

Image Output

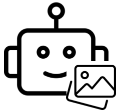
Experiments

2 Groups:

1. Zero-shot
   1. SD_{OPT} (OPT+SD)
   2. SD_{Vicuna} (Vicuna + SD)
   3. GILL
2. Fine-tuning
   1. GILL
   2. GILL with Vicuna

The Department of Justice and Public Safety in the Canadian province of New Brunswick was formed when Premier Brian Gallant restructured government departments in 2016. …. Largely created from the former Department of the Solicitor General, …. The department is headed by a Minister of Justice and Public Safety who also continues to hold the title of Solicitor General of New Brunswick (French: *Ministre de la Sécurité Publique et Solliciteur Général*).

Input Long Text with prompt



Fine-tuned GILL

Image Output

Experiments

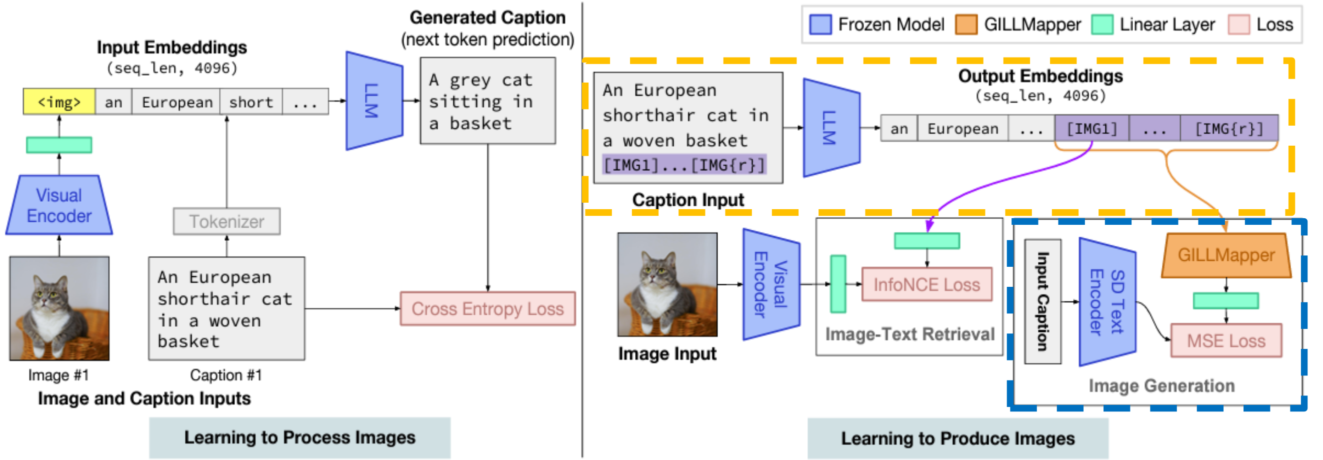


Figure : GILL model architecture overview. It is trained with a captioning loss to learn to process images (left), and losses for image retrieval and image generation to learn to produce images (right).

Input Long Text with prompt

Fine-tuned GILL

Image Output

1. Generate special image tokens
2. Align image token representation with SD text encoder output space

Presentation outline

- Problem Statement & Motivation
- Literature Review
- Dataset
- Experiments
- Metrics
- Results
- Qualitative Examples
- Summary

Metrics

- Challenges:
  - Lack of objective metrics
  - Lack of ground truth
  - Context sensitivity
  - Novelty/Authenticity
  - Diversity
  - Semantic relevance
  - Fidelity/quality
  - Subjectivity/Human bias

Metrics

- Semantic Similarity :
    - CLIP-similarity
    - BLIP-2 similarity
        - S-BERT similarity
        - BERTScore
        - Rouge-1,2,L
- Stylistic Similarity:
    - LPIPS

Metrics

- Semantic Similarity :
  - CLIP-similarity
  - BLIP-2 similarity
    - S-BERT similarity
    - BERTScore
    - Rouge-1,2,L
- Stylistic Similarity:
  - LPIPS

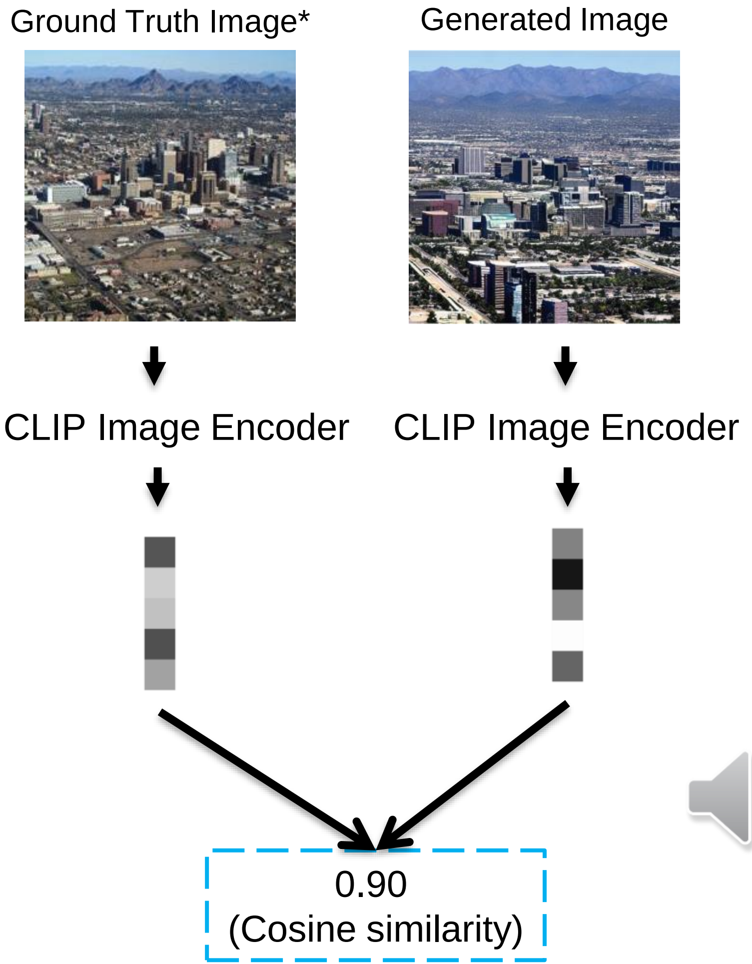


*Image Caption: Downtown Phoenix from an airplane, 2011

Metrics

- ❑ Semantic Similarity :
  - ❑ CLIP-similarity
  - ❑ BLIP-2 similarity
    - ❑ S-BERT similarity
    - ❑ BERTScore
    - ❑ Rouge-1,2,L
- ❑ Stylistic Similarity:
  - ❑ LPIPS

Ground Truth Image*

Generated Image

↓ BLIP-2 ↓ Caption_{gt}

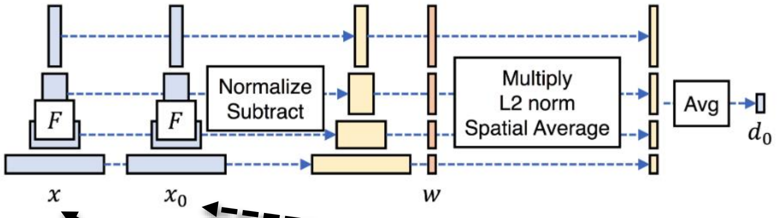
↓ BLIP-2 ↓ Caption_{gen}

0.90
(S-BERT Similarity)

*Image Caption: Downtown Phoenix from an airplane, 2011

Metrics

- ❑ Semantic Similarity :
  - ❑ CLIP-similarity
  - ❑ BLIP-2 similarity
    - ❑ S-BERT similarity
    - ❑ BERTScore
    - ❑ Rouge-1,2,L
- ❑ Stylistic Similarity:
  - ❑ LPIPS



$d_0 = 0.40$

Generated Image, x

Ground Truth Image, x_0*

*Image Caption: 1838 anti-slavery token "Am I not a woman and a sister".

Presentation outline

- ❑ Problem Statement & Motivation
- ❑ Literature Review
- ❑ Dataset
- ❑ Experiments
- ❑ Metrics
- ❑ Results
- ❑ Qualitative Examples
- ❑ Summary

Zero-shot Prompts

Three prompts were evaluated:

- Prompt 1: “Summarize into one sentence that can be used as the caption of a corresponding image”
- Prompt 2: “From this text snippet generate the best caption to describe a relevant image”
- Prompt 3: “Craft a relevant image caption that represents the given text”

Results

Type	Model	CLIP Sim ($\uparrow$)	LPIPS ($\downarrow$)	BLIP-2 Sim_{BERT} ($\uparrow$)	BLIP-2 $\text{Sim}_{\text{S-BERT}}$ ($\uparrow$)	BLIP-2 $\text{Sim}_{\text{ROUGE}}$ ($\uparrow$)
Reference	$\text{SD}_{caption}$	0.6477	0.7151	0.7033	0.4732	0.3462
Zero-shot	SD_{OPT}	0.5599	0.7406	0.6549	0.3364	0.2512
	SD_{Vicuna}	0.5998	0.7314	0.6669	0.3750	0.2692
	GILL	0.5674	0.7359	0.6660	0.3630	0.2624
Fine-tuned	FT-GILL_{OPT}	0.5947	0.7309	0.6798	0.3878	0.2884
	FT-GILL_{Vicuna}	**0.6054**	**0.7241**	**0.6813**	**0.3955**	**0.2925**

Results

Type	Model	CLIP Sim (↑)	LPIPS (↓)	BLIP-2 Sim$_{BERT}$ (↑)	BLIP-2 Sim$_{S\text{-}BERT}$ (↑)	BLIP-2 Sim$_{ROUGE}$ (↑)
Reference	SD$_{caption}$	0.6477	0.7151	0.7033	0.4732	0.3462
Zero-shot	SD$_{OPT}$	0.5599	0.7406	0.6549	0.3364	0.2512
	SD$_{Vicuna}$	0.5998	0.7314	0.6669	0.3750	0.2692
	GILL	0.5674	0.7359	0.6660	0.3630	0.2624
Fine-tuned	FT-GILL$_{OPT}$	0.5947	0.7309	0.6798	0.3878	0.2884
	FT-GILL$_{Vicuna}$	**0.6054**	**0.7241**	**0.6813**	**0.3955**	**0.2925**

- SD$_{vicuna}$ is the best zero-shot model.
- GILL with Vicuna as LLM is the best fine-tuned model.

Presentation outline

- ❑ Problem Statement & Motivation
- ❑ Literature Review
- ❑ Dataset
- ❑ Experiments
- ❑ Metrics
- ❑ Results
- ❑ Qualitative Examples
- ❑ Summary

Qualitative Examples: High CLIP-similarity

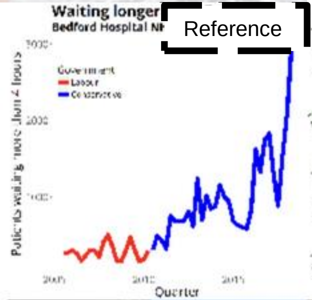
Entrance to the house

Beaumont–Adams percussion revolver

Hector as a tropical depression in the western Pacific Ocean early on August 16

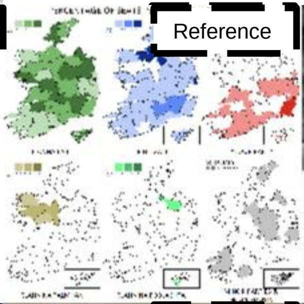
Qualitative Examples: Low CLIP-similarity

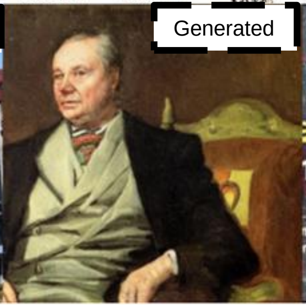
Four-hour target in the emergency department quarterly figures from NHS England Data from https://www.england.nhs.uk/statistics/statistical-work-areas/ae-waiting-times-and-activity/



Bedford Hospital / Performance

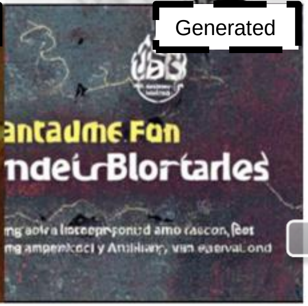
Percentage of seats gained by each of the five biggest parties, and number of seats gained by smaller parties and independents.





1951 Irish general election

Albert Finney won for his portrayal of Edward L. Masry in Erin Brockovich (2000)



Screen Actors Guild Award for Outstanding Performance by a Male Actor in a Supporting Role / Winners and nominees

Qualitative Examples : Low LPIPS

Eisenhower dollar obverse design used from 1971-1978. This particular coin is the silver version of the coin minted in 1974 at the San Francisco mint and graded MS67 by PCGS. Note the small "S" mint mark below the bust of Eisenhower but above the date digits "74"

Commemorative coin - obverse

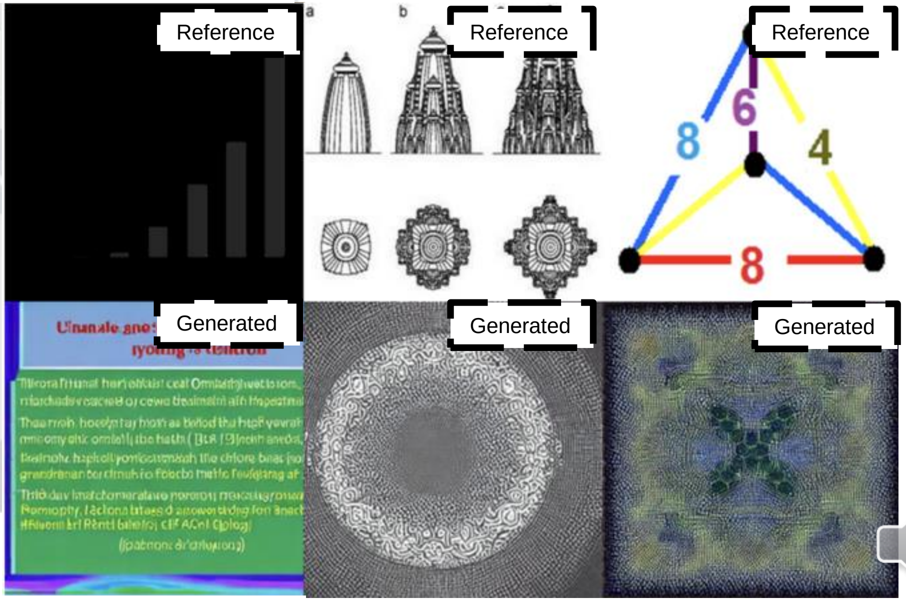
Official portrait of Jessica Morden MP

Qualitative Examples: High LPIPS

Average annual number of DF cases and DHF cases reported to WHO', 'Hindu temples feature self-similar, fractal-like structures, where parts resemble the whole.[1]

Hindu temples feature self-similar, fractal-like structures, where parts resemble the whole.[1]

Paracompact uniform honeycomb



Dengue fever outbreaks

Fractal art

Paracompact uniform honeycombs / Cyclic graphs / [(4,4,4,3)] family

Qualitative Examples: Comparison

Example 1: The Shah Jahan Mosque, also known as the Jamia Masjid of Thatta, is a 17th-century building that serves as the central mosque for the city of Thatta, in the Pakistani province of Sindh. The mosque is considered to have the most elaborate display of tile work in South Asia, and is also notable for its geometric brick work - a decorative element that is unusual for Mughal-period mosques. It was built during the reign of Mughal emperor Shah Jahan, who bestowed it to the city as a token of gratitude, and is heavily influenced by Central Asian architecture - a reflection of Shah Jahan's campaigns near Samarkand shortly before the mosque was designed.

Ground Truth

FT-GILL$_{Vicuna}$

Qualitative Examples: Comparison

Example 5: Both mining and logging create similar secondary deforestation through road construction. Specifically, logging companies construct new roads into previously inaccessible forest areas which facilitates the conversion of logged forests by into agricultural land. This has led to the immigration of landless farmers, in particular from eastern savanna regions, to enter primary forest areas through logging roads. Incoming farmers cause extensive land degradation in converting the natural forest into farmlands. Further, it has been suggested that increases in returns can lead to substantial increase in farm sizes and shortening of the fallow period, which in turn eventually leads to large-scale and severe natural forest area destruction.

Ground Truth

FT-GILL$_{Vicuna}$

Presentation outline

- Problem Statement & Motivation
- Literature Review
- Dataset
- Experiments
- Metrics
- Results
- Qualitative Examples
- Summary

Summary

1. Investigated the task of contextual image generation from long-form text from the perspective of LLMs and TIMs.
2. Compared zero-shot prompting and supervised fine-tuning approaches for this task.
3. Introduced the novel BLIP-2 similarity metric to evaluate the semantic correctness of generated images.
4. Established baselines and provided insights into the strengths and limitations of existing models for image generation from long-form text.

References

[1] Xu, Peng, et al. "Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models." *arXiv preprint arXiv:2306.09265* (2023).
[2] Koh, Jing Yu, Daniel Fried, and Ruslan Salakhutdinov. "Generating images with multimodal language models." *arXiv preprint arXiv:2305.17216* (2023).
[3] Aghajanyan, Armen, et al. "Cm3: A causal masked multimodal model of the internet." *arXiv preprint arXiv:2201.07520* (2022).
[4] Srinivasan, Krishna, et al. "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning." *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021.

Thanks

for your time!!!