ConZIC: Controllable Zero-shot Image Captioning by Sampling-Based Polishing

Zequn Zeng*, Hao Zhang*, Ruiying Lu, Dongsheng Wang, Bo Chen†
National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, China
{zzequn99, zhanghao_xidian}@163.com, bchen@mail.xidian.edu.cn

Zhengjue Wang
State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, 710071, China
zhengjuewang@163.com

THU-PM-273

XIDIAN UNIVERSITY 1931

SigMa
Signal Processing & Machine Learning

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Traditional image captioning

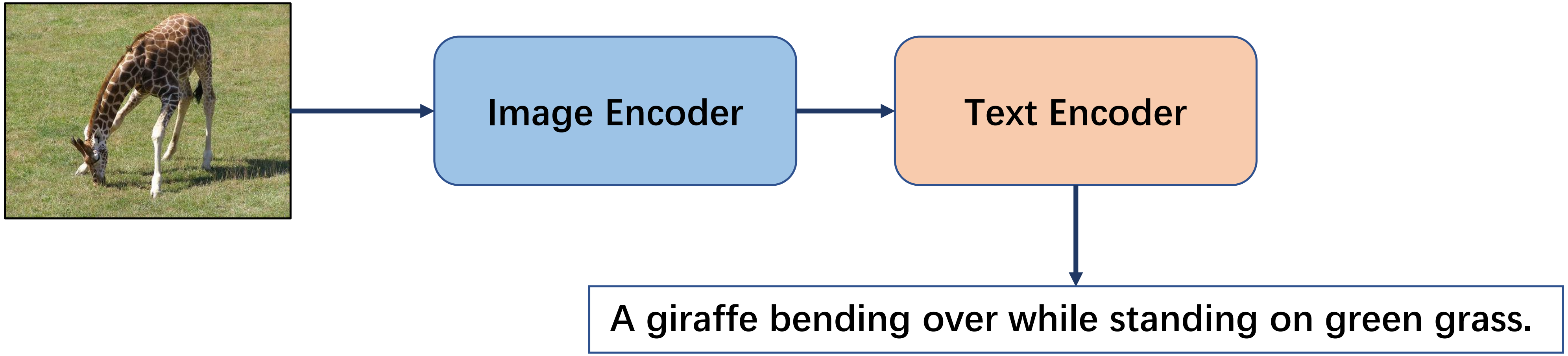


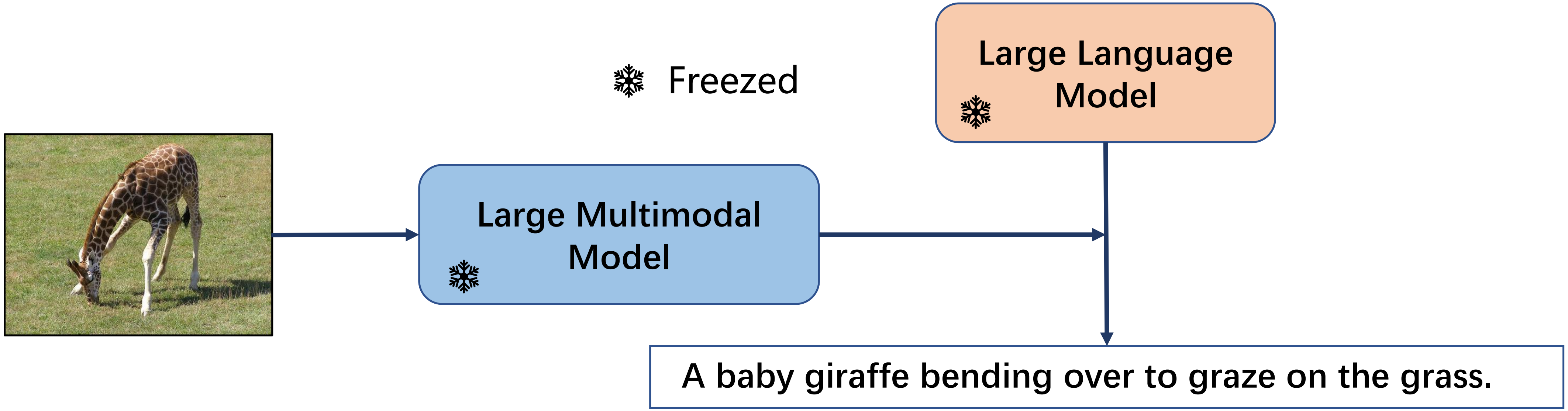
Image Encoder

Text Encoder

A giraffe bending over while standing on green grass.

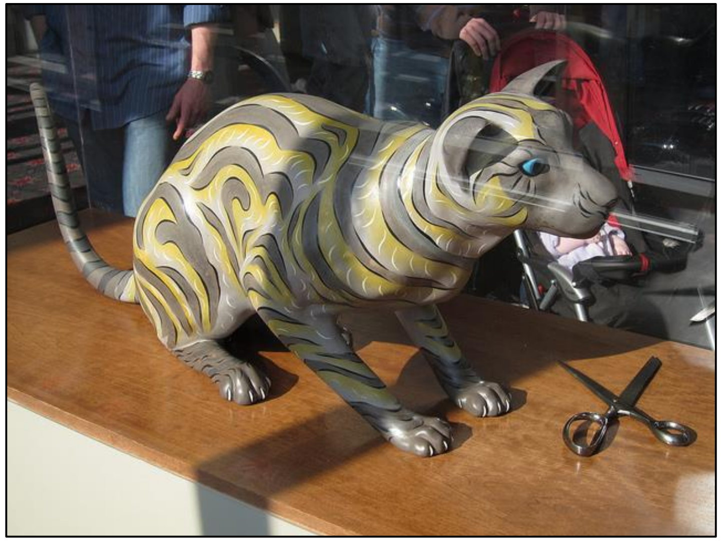
- **Supervised training** with curated **image-text pairs**
- **Closed-set:** testing and training **I.I.D**
- **Limitation:** lack of **robustness** and **diversity**

Zero-shot image captioning



- **Zero-shot** without training data
- **Open-set：Extensive visual concepts** in large multimodal model, e.g., **CLIP**
- **Knowledge** in Large pretrained model lead to **high robustness**

Motivation



Autoregressive generation:

A dog replica.
A dog sculpture.
A dog statue.
A dog sculpture created in London's Museum of Modern Art.
A dog sculpture created in London's Museum of Modern Art in the early 2000s.

Mode collapse

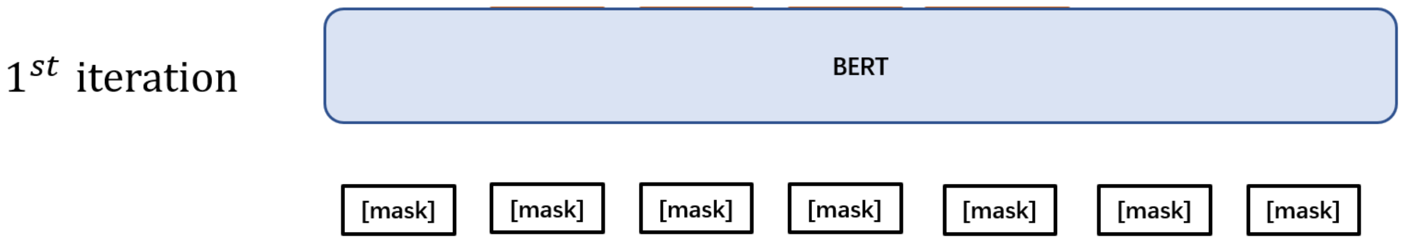
Our non-autoregressive generation:

A striped 3d pet model-sized grey lab tiger displayed.
A grey metallic 3d model exhibiting a striped pet tiger.
A tiger sculpture painted on a statue display shown throughout campus.
A silver painted animal in striped yellow within window displays.
A silver striped tiger model depicted on window shopping display.

- **Mode collapse: similar syntactic patterns** and **low diversity**
- **Less flexible**: left-to-right generation and no chance to **modify** generated word
- **Lack of controllability:** sentiment, style, personality

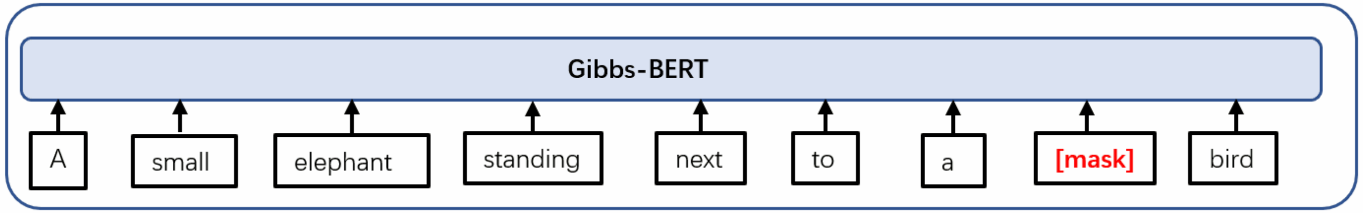
Non-autoregressive LM: Gibbs-BERT

Gibbs-BERT

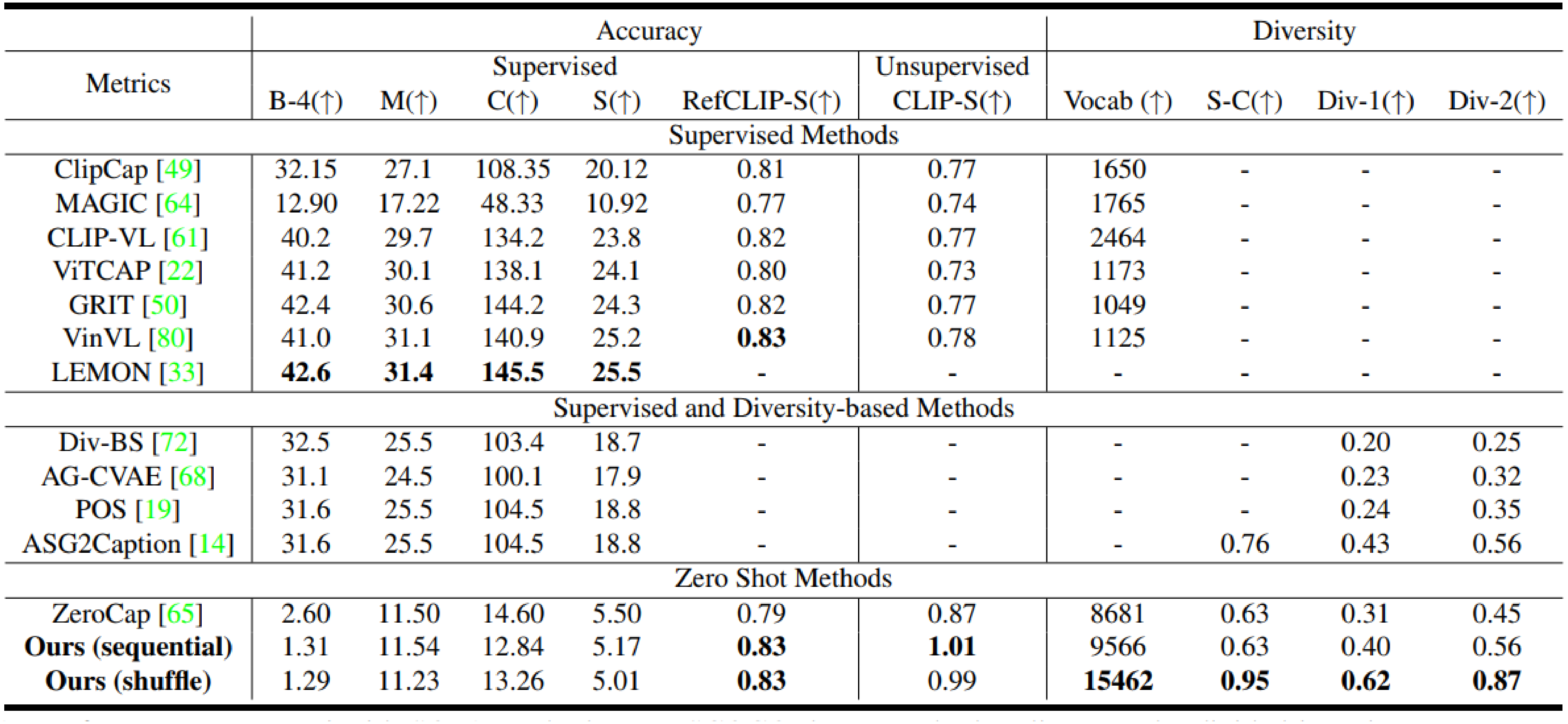


Replace word in arbitrary order

Proposed method

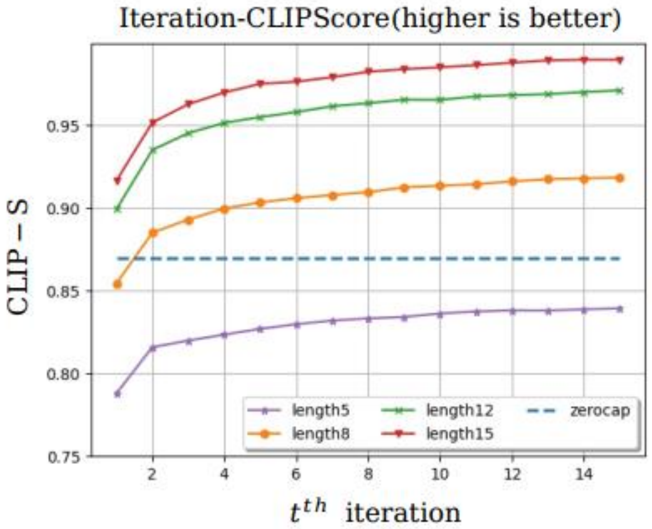


Performance on MSCOCO

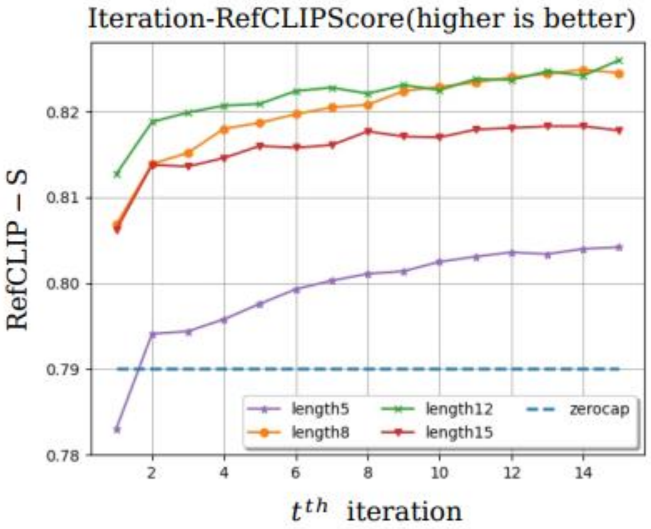


Metrics	Accuracy						Diversity			
	Supervised					Unsupervised				
	B-4(↑)	M(↑)	C(↑)	S(↑)	RefCLIP-S(↑)	CLIP-S(↑)	Vocab (↑)	S-C(↑)	Div-1(↑)	Div-2(↑)
Supervised Methods										
ClipCap [49]	32.15	27.1	108.35	20.12	0.81	0.77	1650	-	-	-
MAGIC [64]	12.90	17.22	48.33	10.92	0.77	0.74	1765	-	-	-
CLIP-VL [61]	40.2	29.7	134.2	23.8	0.82	0.77	2464	-	-	-
ViTCAP [22]	41.2	30.1	138.1	24.1	0.80	0.73	1173	-	-	-
GRIT [50]	42.4	30.6	144.2	24.3	0.82	0.77	1049	-	-	-
VinVL [80]	41.0	31.1	140.9	25.2	**0.83**	0.78	1125	-	-	-
LEMON [33]	**42.6**	**31.4**	**145.5**	**25.5**	-	-	-	-	-	-
Supervised and Diversity-based Methods										
Div-BS [72]	32.5	25.5	103.4	18.7	-	-	-	-	0.20	0.25
AG-CVAE [68]	31.1	24.5	100.1	17.9	-	-	-	-	0.23	0.32
POS [19]	31.6	25.5	104.5	18.8	-	-	-	-	0.24	0.35
ASG2Caption [14]	31.6	25.5	104.5	18.8	-	-	-	0.76	0.43	0.56
Zero Shot Methods										
ZeroCap [65]	2.60	11.50	14.60	5.50	0.79	0.87	8681	0.63	0.31	0.45
Ours (sequential)	1.31	11.54	12.84	5.17	**0.83**	**1.01**	9566	0.63	0.40	0.56
Ours (shuffle)	1.29	11.23	13.26	5.01	**0.83**	0.99	**15462**	**0.95**	**0.62**	**0.87**

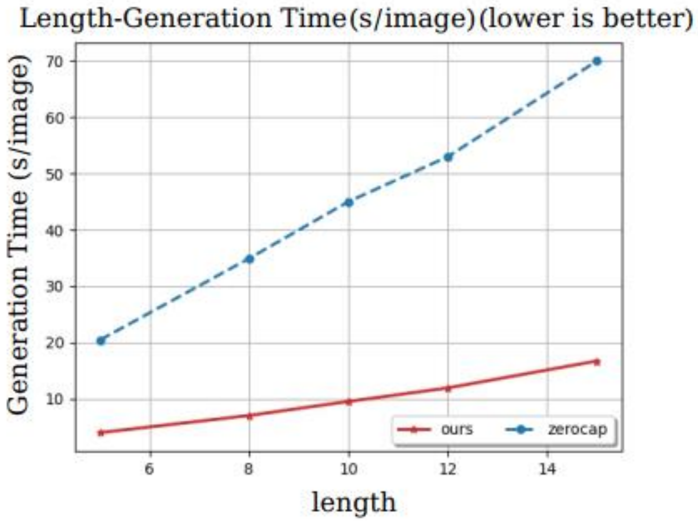
Iterative curve and time-consuming



(a) Comparision on CLIPScore.

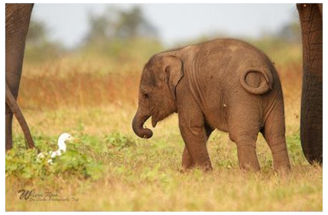


(b) Comparision on RefCLIPScore.



(c) Comparision on Time-consuming.

Controllable generation and Comparison with SOTA



+ ☺ **Positive**

or

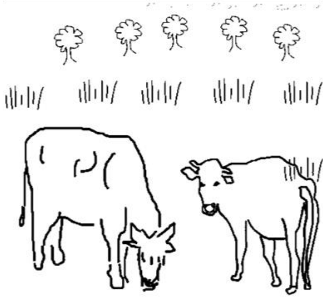
☹ **Negative**

Positive:

1. A very ***cute cheerful*** white ***bird accompanies*** a ***happy*** tiny elephant.
2. A cute ***little*** white ***duck enjoys amazed*** chatting with an elephant.
3. A white ***hen*** and an extremely ***small beautiful*** elephant ***play*** *happily* on ground.
4. A ***healthy*** elephant ***enthusiastically admires*** an ***awesome adorable southern bird***.
5. A ***gorgeous*** elephant ***walks*** beside a white ***goose*** with miniature ***smiling***.

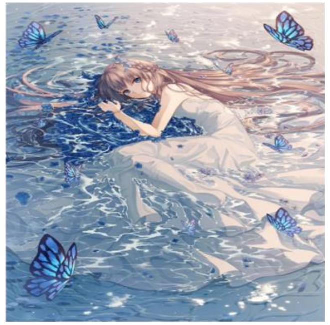
Negative:

1. A ***badly*** elephant ***stars*** at a ***scared small bird***.
2. A scared little bird is ***afraid*** that the ***vicious*** elephant would eat it.
3. Image of a ***sad libelous chicken moping*** alongside a small ***lonely*** elephant is shown.
4. A ***stray worried*** white ***duck meets*** a ***stealthy hungry*** elephant.
5. A ***brown solitary*** elephant ***roams*** with an ***lonely*** white ***sparrow*** nearby.



GRIT: *A drawing of graffiti on a wall.*
ViTCap: *A picture of a sheep and a cow.*
CLIPCap: *A black and white photo of a group of cows.*
ZeroCap: *Image of a cows drawing.*

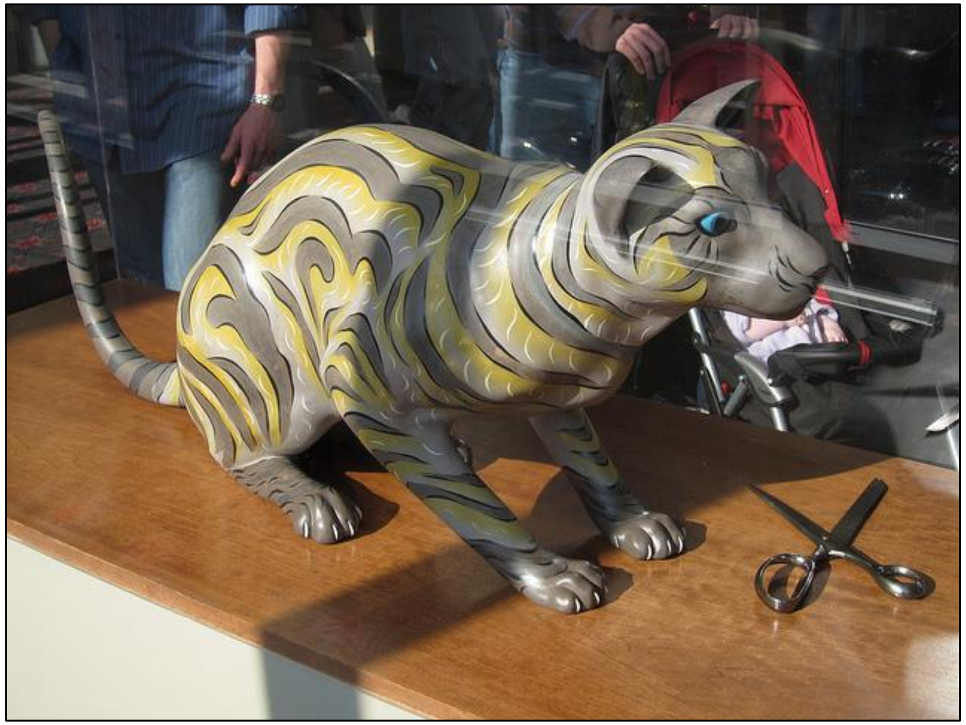
Ours: ***Two cows face each other on a pasture with various flowers in sequence.***



GRIT: *A woman sitting on a fountain in the sea.*
ViTCap: *A painting of a woman laying on a bed.*
CLIPCap: *A painting of a woman in on a surfboard.*
ZeroCap: *Image of a girl sleeping in the sea.*

Ours: ***A painting of the princess submerged in delicate poses with water background.***

Diverse generation comparison



ZeroCap(beam 5):

A dog replica.

A dog sculpture.

A dog statue.

A dog sculpture created in London's Museum of Modern Art.

A dog sculpture created in London's Museum of Modern Art in the early 2000s.

Ours(shuffle):

Order: 7, 3, 2, 8, 5, 6, 9, 4, 0, 1

Cap: **A striped 3d pet model-sized grey lab tiger** displayed.

Order: 7, 8, 1, 5, 3, 4, 2, 0, 9, 6

Cap: **A grey metallic 3d model** exhibiting a striped pet tiger.

Order: 6, 8, 9, 7, 5, 3, 0, 4, 1, 2

Cap: **A tiger sculpture** painted on a statue display shown throughout campus.

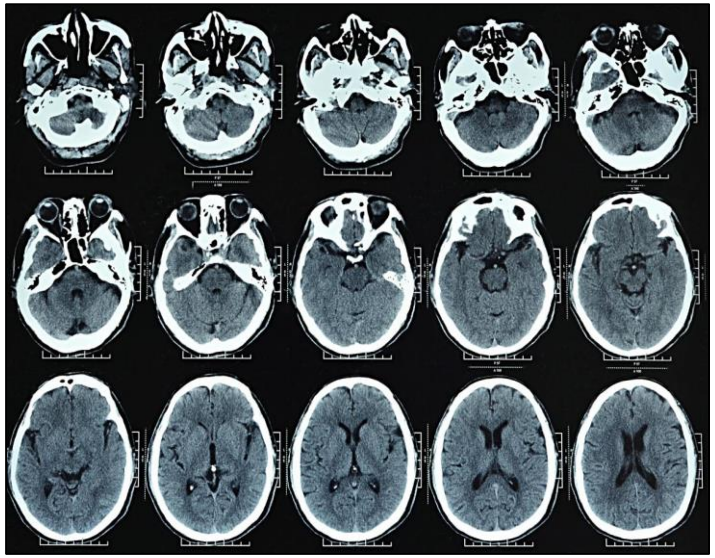
Order: 5, 9, 3, 4, 6, 7, 2, 8, 1, 0

Cap: **A silver painted animal in striped yellow** within window displays.

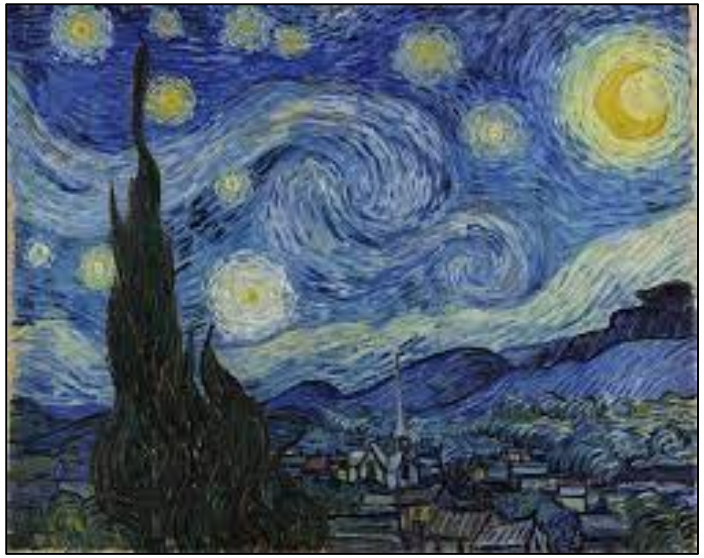
Order: 1, 5, 6, 0, 9, 4, 7, 2, 8, 3

Cap: **A silver striped tiger model** depicted on window shopping display.

World knowledge

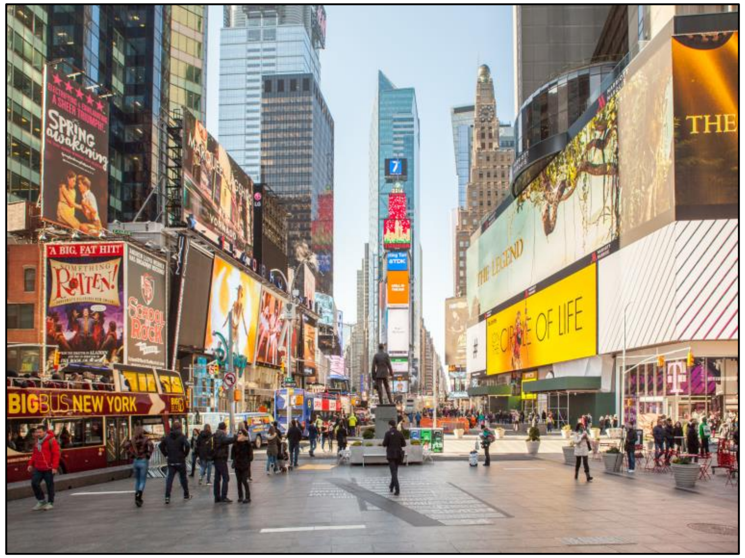


GRIT: A series of clocks on top of a screen
CLIPCap: A picture of a skull and crossbones with a lot of wires.
ViTCap:a close up of a cake with a picture of a dog.
ZeroCap: A patient submitted to the neurologist's office.
Ours:
A complete **CT medical photo** with **brain** samples.
A **CT scene** multiple frames displaying a diagnosis.
A stacked sheet displaying signs of brain damaged.
A composite present multiple images featuring frontal trauma.



GRIT: A painting of a painting with a tree in the background
CLIPCap: The night sky over the city.
ViTCap:A painting of a bird on a table with a bird on it.
ZeroCap: A night with Vincent.

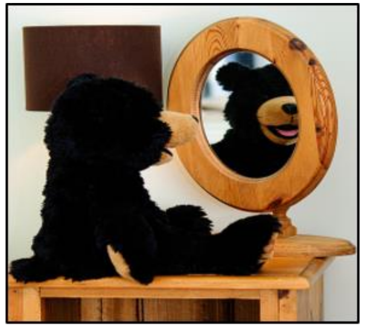
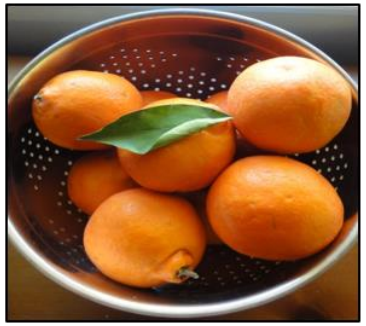
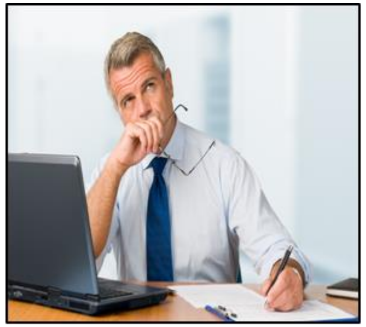
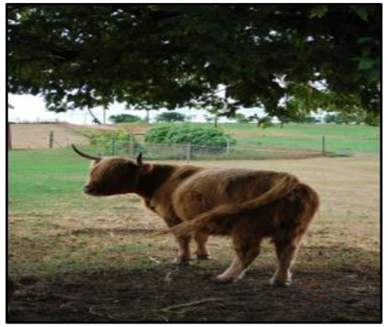
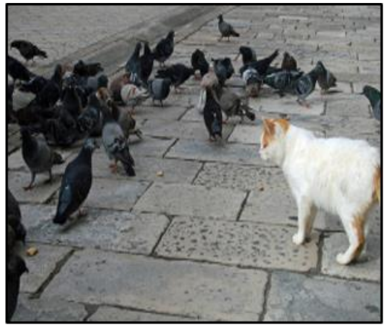
Ours:
A famous **Gogh painting** after streaming moonlight over all the grand structures.
A view despite a nocturnal sky within **famous mainstream artworks**.
A nighttime sky can appear in drawings and oil paintings.

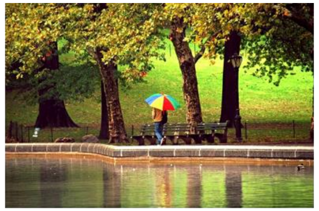
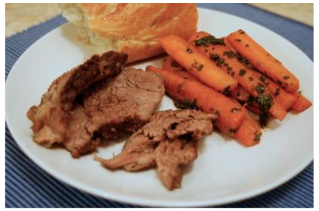
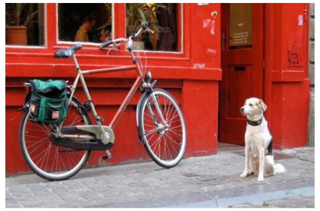


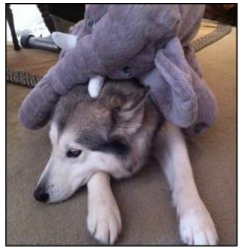
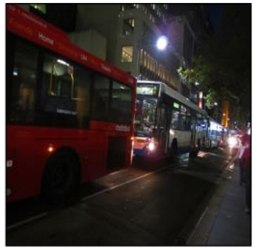
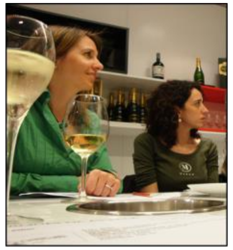
GRIT: A busy city street with lots of people walking on.
CLIPCap:A busy city street with people crossing it.
ViTCap: a city street with people walking and a bus.
ZeroCap:A billboard in the middle of of of a busy intersection.

Ours:
A new york **time square**.
A busy billboard covered **time square** with scenery.
A landscape masking time square depicting a vibrant morning.
A city featuring **yellow billboard and advertisements** on google outdoor.

More controllable generation

sentence length control					
3~5	A stuffed black bear.	A fruit dish.	A calm businessman concentrating hard.	A blond farm cow.	A central animal courtyard.
7~9	A bear toy named Cooper admiring himself.	A fruit dish in tin color offering sweet orange.	A financial administrator watching financial statements online.	A farm buffalo around metal enclosure and foliage.	A cornered cat shown against numerous pigeons.
11~13	A stuffed teddy dark bear smiling with yoga pose in a mirror.	A photo showing Osaka orange fruits appearing in a stainless steel pot.	A man distracted thinking business report with neatly trimmed white hair.	A village animal cow shows in tree ferns background and fences.	A mute cat meeting numerous birds and pigeons in a Greek square.

Style Control			
Factual	A lone teenager walking through a park bench over water with an umbrella.	A beef protein meal includes carrots and a roasted duck sandwich.	A white Irish dog and a bicycle outside a French red fabric shop.
Positive	A park walker enjoying rain and pleasant colors in a gorgeous atmosphere.	A sandwich recipe includes wonderful carrots and a perfect beef roast.	A french dog enjoying popularity in a gorgeous red-painted bicycle store.
Negative	A park scenery depicts an lonely depressed person in yellow raincoat.	A fake roast meat beef dinner provides faux carrot sandwiches.	A generic red bicycle beside an unhappy Irish tourist dog.

Parts-of-speech (POS) control			
POS	*DET ADJ/NOUN NOUN VERB VERB ADV ADP DET ADJ/NOUN NOUN NOUN.*		
Without control	A gray dog embracing a familiar sad purple elephant with shark clothing.	A darkened London lane with a red and gray southbound bus pair.	A female group members in a Portuguese bar counter tasting sparkling wines.
POS control	The grey dog is embraced unexpectedly by a harmless purple elephant.	The electric buses can be seen on a busy Sydney night.	Some female guiders are encountered together during Brazilian wine bar classes.

Thanks for watching!