MeaCap: Memory-Augmented Zero-shot Image Captioning

Zequn Zeng[*], Yan Xie[*], Hao Zhang[†], Chiyu Chen, Bo Chen[†]
National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, China
{zzequn99, yanxie0904, zhanghao_xidian}@163.com, {chenchiyu, bchen}@mail.xidian.edu.cn

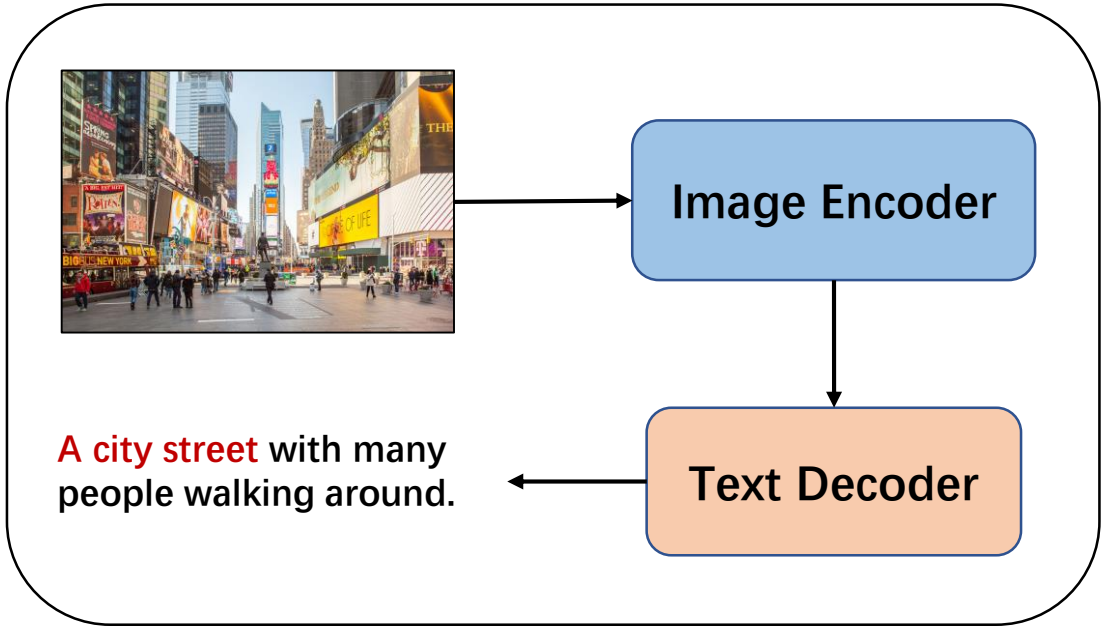
Zhengjue Wang
State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, 710071, China
wangzhengjue@xidian.edu.cn

Arch 4A-E Poster #430

Signal Processing & Machine Learning

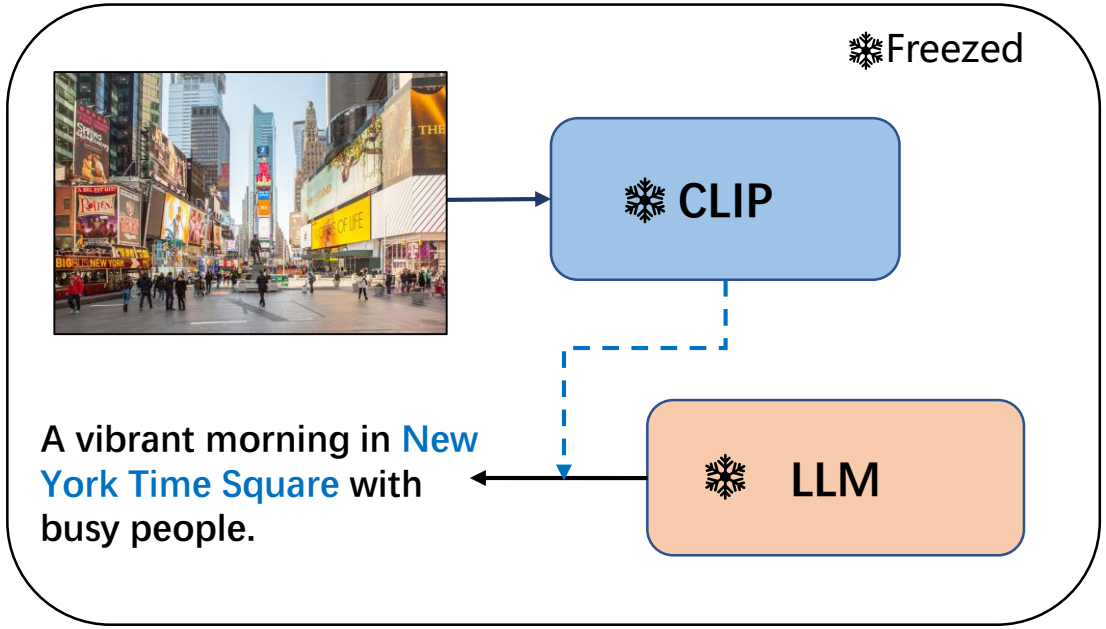
Computer Vision and Pattern Recognition Conference
Seattle | June 17-21, 2024

Traditional image captioning



- **Supervised training:** require clean image-text pairs
- **Closed-set:** impossible to describe novel objects
- **General captioning**: lack of world knowledge and diversity

Zero-shot image captioning

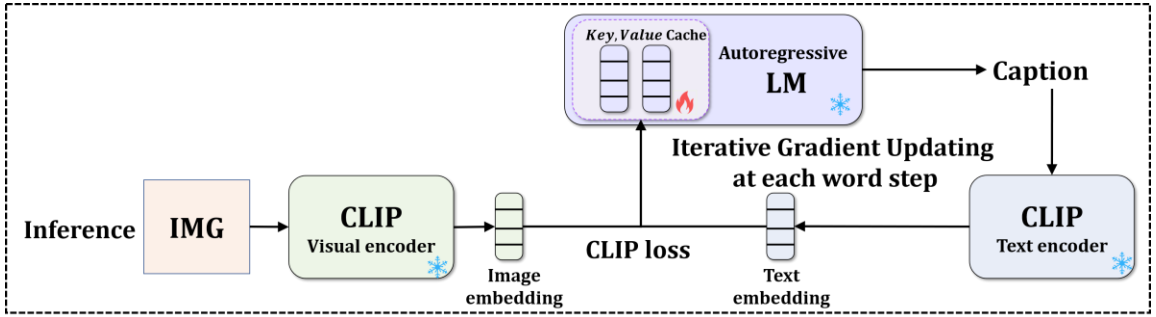


- **Training-free or Text-only-training**
- **Open-set:** cover extensive visual concepts
- **Knowledge-enhanced**: abundant world knowledge in CLIP and LLM

Zero-shot image captioning

Training-free methods:

- Require no data
- Frozen CLIP + Frozen LLM

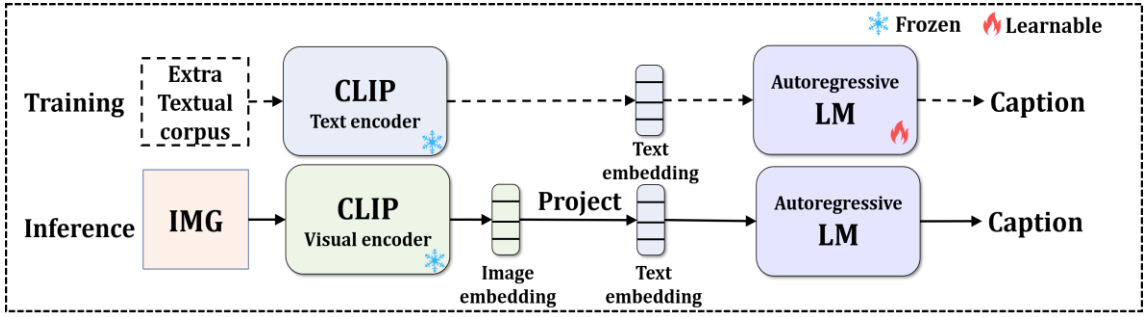


Text-only-training	CLIP Score	BLIP-2 Score
MAGIC: A plate topped with cake and spoon.	0.76	0.89
DeCap: A piece of cake on a white plate with a spoon.	0.77	0.87
ViECap: Cake with white frosting on a white plate on a table.	0.75	0.73
$MeaCap_{ToT}$: *concepts*: [slice lemon pie, serving plate] *caption*: A slice of lemon pie with spoon on serving plate on table.	0.83	0.83
Training-free		
ZeroCap: A large dessert eaten in the 2016 New Hampshire State Hotel.	0.87	0.75
ConZIC: A butter pie served at the famous Mary Teresa restaurant.	1.00	0.77
$MeaCap_{TF}$: *concepts*: [slice lemon pie, serving plate] *caption*: A slice of lemon pie with a spoon on a serving plate.	0.84	0.82

Hallucination generation

Text-only-training methods:

- Require textual caption corpus
- Frozen CLIP + Learnable LLM

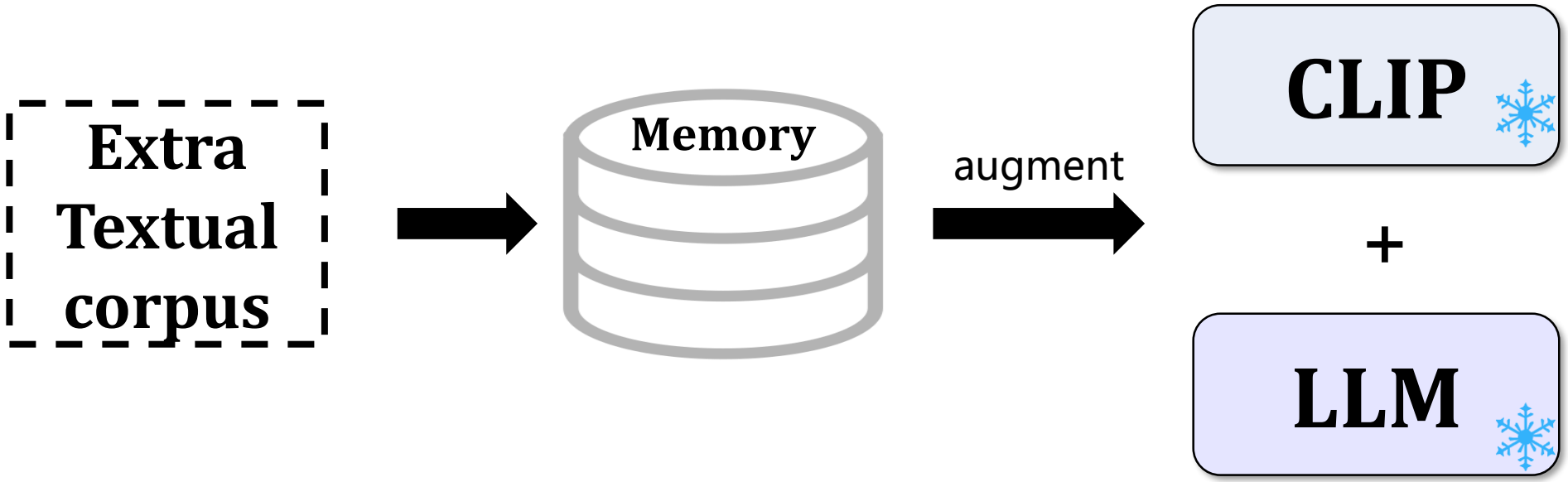


Text-only-training	CLIP Score	BLIP-2 Score
MAGIC: A red and white locomotive is being docked.	0.32	0.30
DeCap: A person that is on the ground and is holding his device.	0.51	0.22
ViECap: Before and after shots of a man in a suit and tie.	0.42	0.31
$MeaCap_{ToT}$: *concepts*: [spiderman] *caption*: A picture of a spiderman comics.	0.68	0.65
Training-free		
ZeroCap: Image of a Web Hero.	0.74	0.27
ConZIC: A very attractive spiderman typical marvel definition.	0.82	0.59
$MeaCap_{TF}$: *concepts*: [spiderman] *caption*: A comic book superhero called spiderman.	0.77	0.68

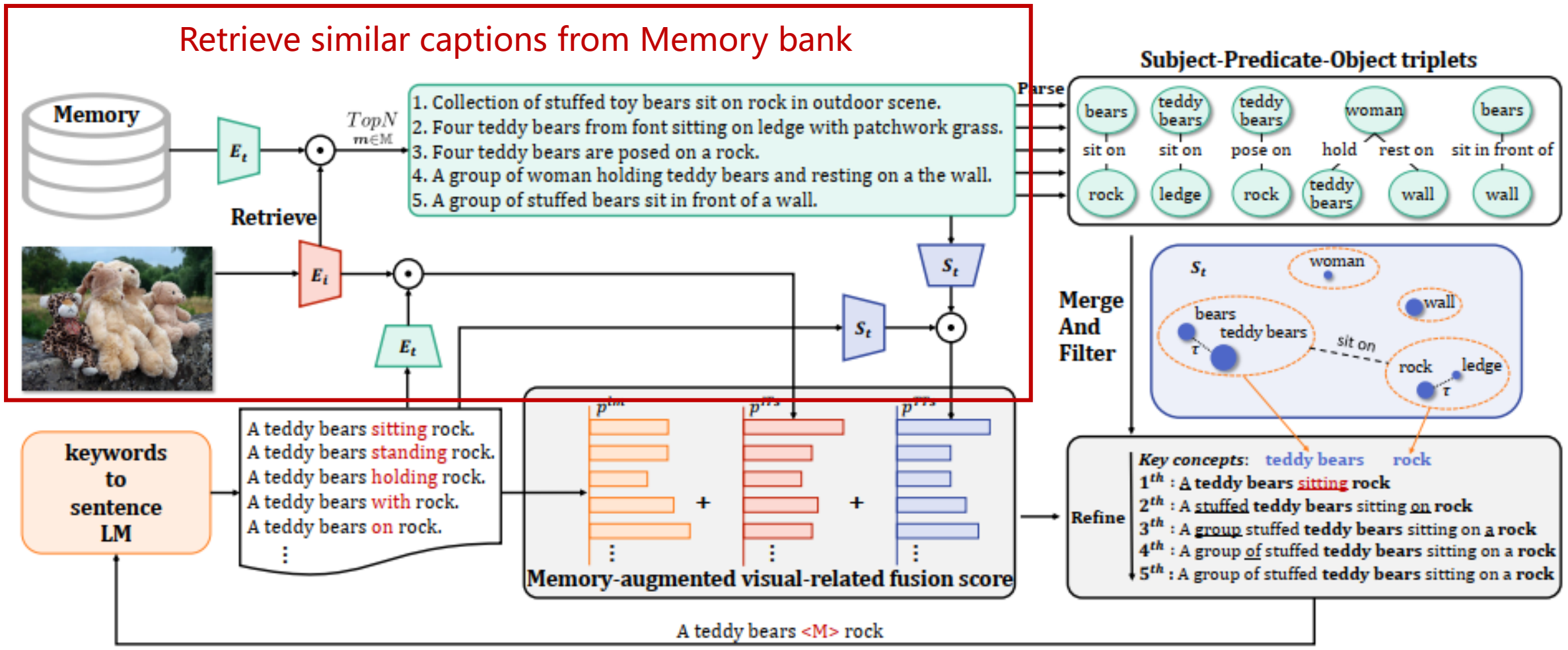
Knowledge forgetting

Motivation

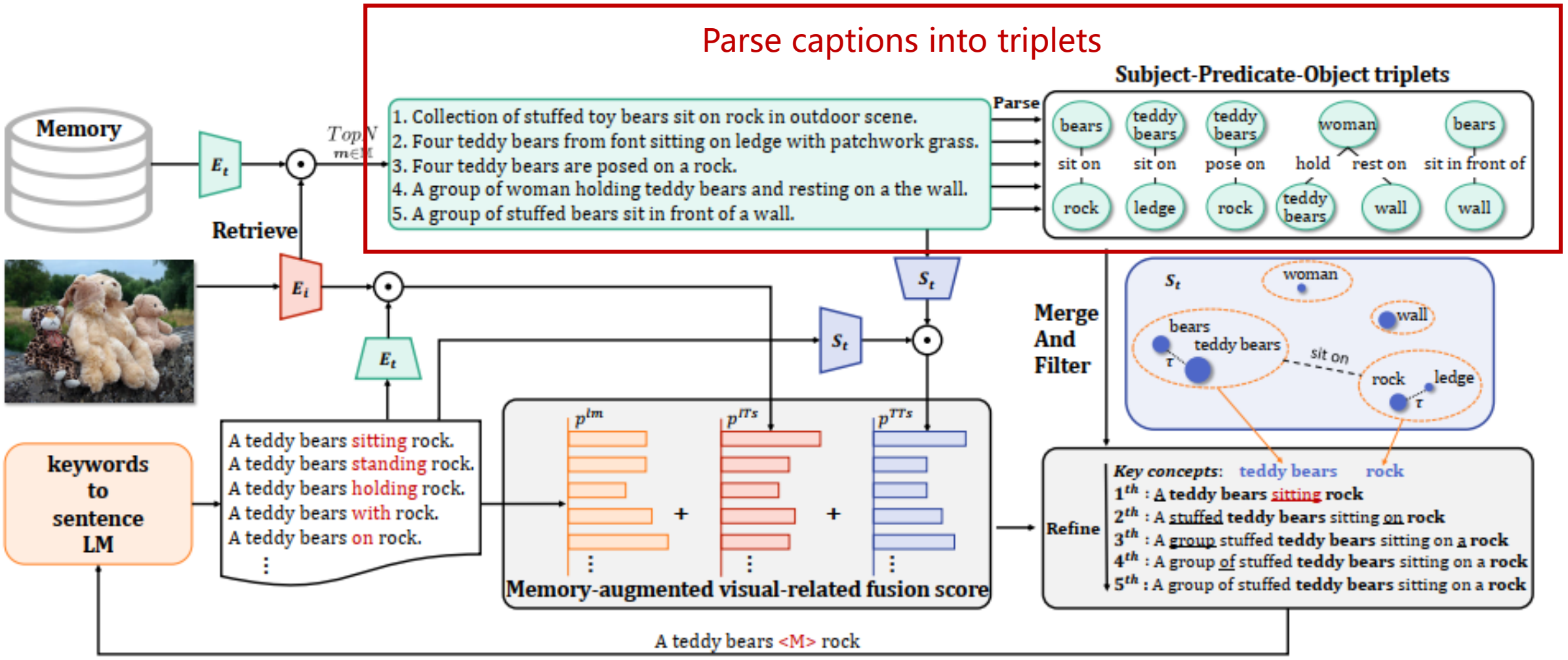
- To maintain good generalization ability to images in the wild and to get rid of unreasonable imagination
- To provide an alternative scheme to use captioning corpus rather than using it to train the LM.



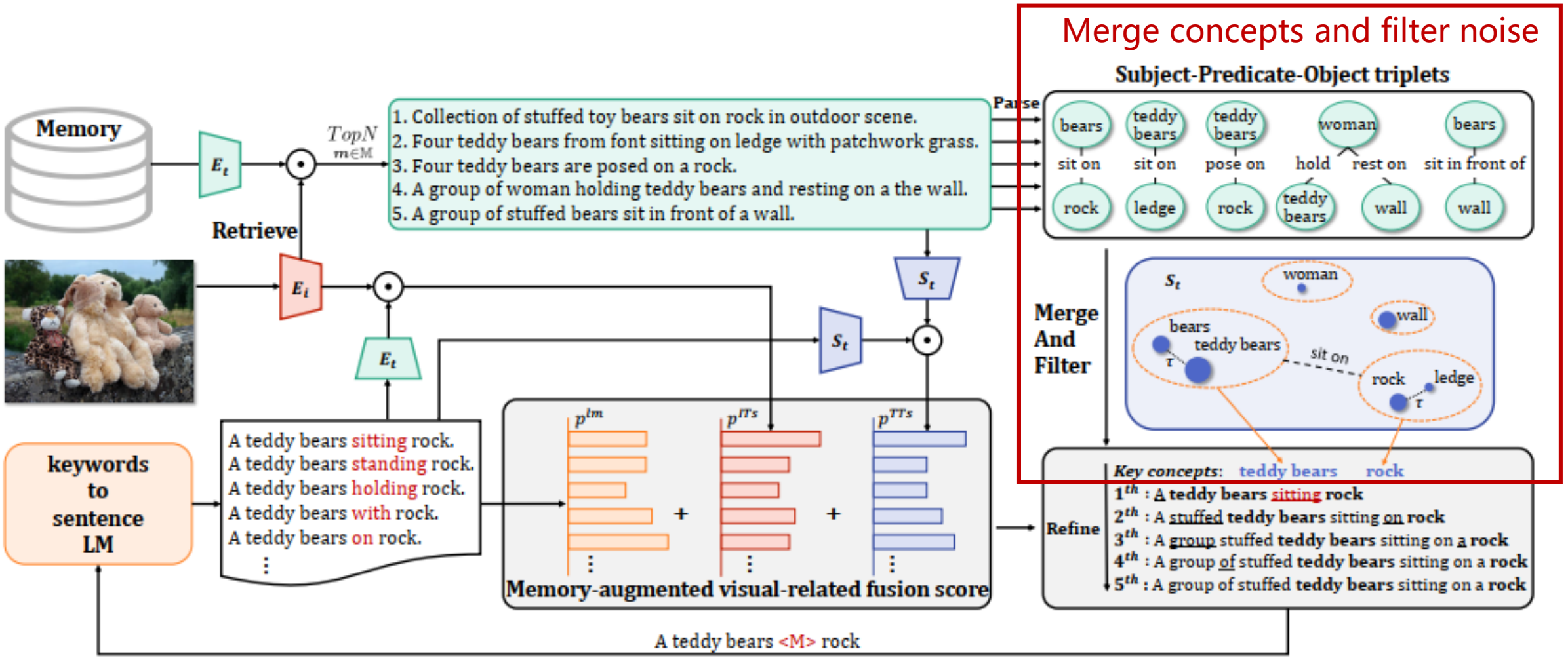
Method


Retrieve similar captions from Memory bank
Memory
E_t
$TopN_{m \in M}$
1. Collection of stuffed toy bears sit on rock in outdoor scene.
2. Four teddy bears from font sitting on ledge with patchwork grass.
3. Four teddy bears are posed on a rock.
4. A group of woman holding teddy bears and resting on a the wall.
5. A group of stuffed bears sit in front of a wall.
Retrieve
E_i
E_t
S_t
S_t
Parse
Subject-Predicate-Object triplets
bears
teddy bears
teddy bears
woman
bears
sit on
sit on
pose on
hold
rest on
sit in front of
rock
ledge
rock
teddy bears
wall
wall
Merge And Filter
S_t
woman
wall
bears
teddy bears
sit on
rock
ledge
keywords to sentence LM
A teddy bears sitting rock.
A teddy bears standing rock.
A teddy bears holding rock.
A teddy bears with rock.
A teddy bears on rock.
Memory-augmented visual-related fusion score
Refine
Key concepts: teddy bears rock
1th : A teddy bears sitting rock
2th : A stuffed teddy bears sitting on rock
3th : A group stuffed teddy bears sitting on a rock
4th : A group of stuffed teddy bears sitting on a rock
5th : A group of stuffed teddy bears sitting on a rock
A teddy bears <M> rock

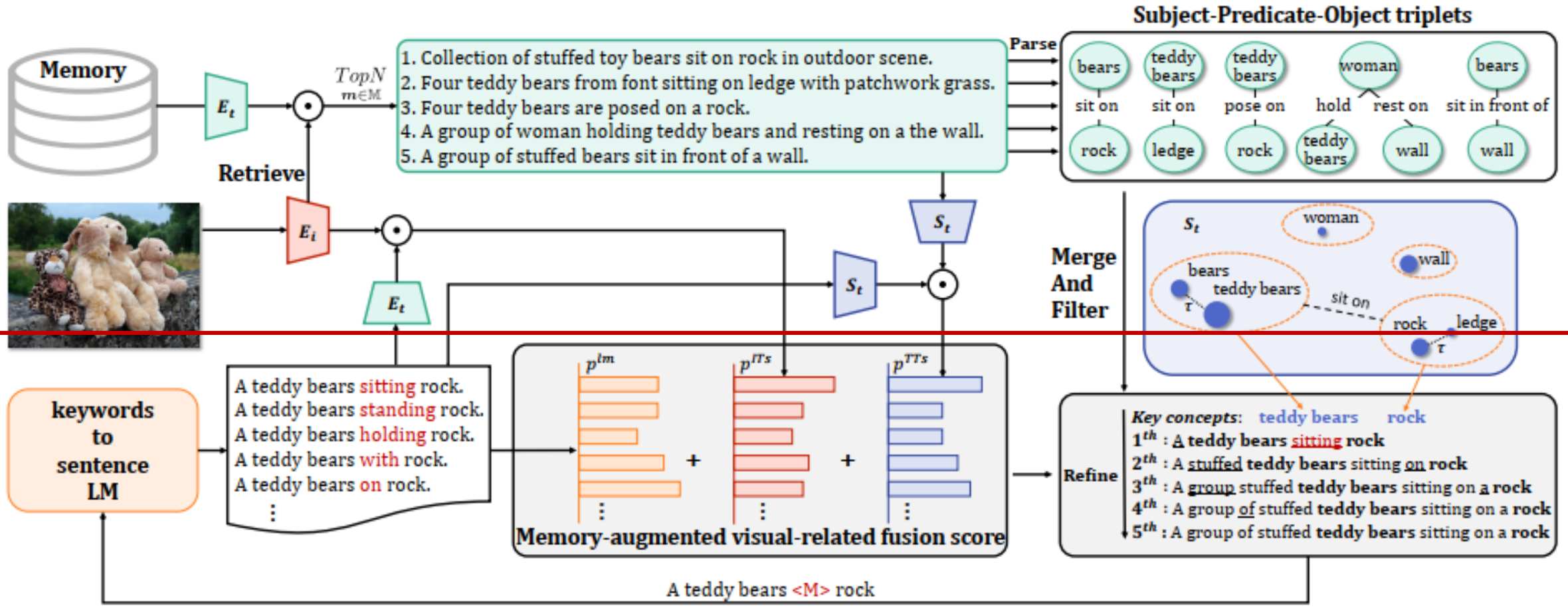
Method


Parse captions into triplets
Subject-Predicate-Object triplets
Memory
Retrieve
Top N
1. Collection of stuffed toy bears sit on rock in outdoor scene.
2. Four teddy bears from font sitting on ledge with patchwork grass.
3. Four teddy bears are posed on a rock.
4. A group of woman holding teddy bears and resting on a the wall.
5. A group of stuffed bears sit in front of a wall.
Parse
bears
teddy bears
teddy bears
woman
bears
sit on
sit on
pose on
hold
rest on
sit in front of
rock
ledge
rock
teddy bears
wall
wall
keywords to sentence LM
A teddy bears sitting rock.
A teddy bears standing rock.
A teddy bears holding rock.
A teddy bears with rock.
A teddy bears on rock.
Memory-augmented visual-related fusion score
Merge And Filter
woman
wall
bears
teddy bears
sit on
rock
ledge
Refine
Key concepts: teddy bears rock
1th : A teddy bears sitting rock
2th : A stuffed teddy bears sitting on rock
3th : A group stuffed teddy bears sitting on a rock
4th : A group of stuffed teddy bears sitting on a rock
5th : A group of stuffed teddy bears sitting on a rock
A teddy bears <M> rock

Method


Merge concepts and filter noise
Memory
Retrieve
TopN
1. Collection of stuffed toy bears sit on rock in outdoor scene.
2. Four teddy bears from font sitting on ledge with patchwork grass.
3. Four teddy bears are posed on a rock.
4. A group of woman holding teddy bears and resting on a the wall.
5. A group of stuffed bears sit in front of a wall.
Parse
Subject-Predicate-Object triplets
bears
teddy bears
teddy bears
woman
bears
sit on
sit on
pose on
hold
rest on
sit in front of
rock
ledge
rock
teddy bears
wall
wall
Merge And Filter
woman
wall
bears
teddy bears
sit on
rock
ledge
keywords to sentence LM
A teddy bears sitting rock.
A teddy bears standing rock.
A teddy bears holding rock.
A teddy bears with rock.
A teddy bears on rock.
Memory-augmented visual-related fusion score
Refine
Key concepts: teddy bears rock
1th : A teddy bears sitting rock
2th : A stuffed teddy bears sitting on rock
3th : A group stuffed teddy bears sitting on a rock
4th : A group of stuffed teddy bears sitting on a rock
5th : A group of stuffed teddy bears sitting on a rock
A teddy bears <M> rock

Method



Iterative insert/replace words based on retrieved concepts

Zero-shot captioning results on MSCOCO and NoCaps

Our training-free version compared to previous training-free methods

Methods	Text Corpus		MSCOCO						NoCap val (CIDEr)			
	Training	Memory	B@4	M	C	S	CLIP-S	BLIP2-S	In	Near	Out	Overall
ZeroCap [55]	✗	✗	2.6	11.5	14.6	5.5	0.87	0.70	13.3	14.9	19.7	16.6
Tewel *et al.* [54]	✗	✗	2.2	12.7	17.2	7.3	0.74	0.68	13.7	15.8	18.3	16.9
ConZIC [65]	✗	✗	1.3	11.2	13.3	5.0	**1.00**	0.76	15.4	16.0	20.3	17.5
CLIPRe [30]	✗	CC3M	4.6	13.3	25.6	9.2	0.84	0.70	23.3	26.8	36.5	28.2
DeCap [30]	CC3M	CC3M	8.8	16.0	42.1	10.9	0.76	-	34.8	37.7	49.9	39.7
MeaCap$_{TF}$	✗	CC3M	7.1	16.6	42.5	11.8	0.84	**0.81**	35.3	39.0	45.1	40.2
MeaCap$_{ToT}$	CC3M	CC3M	**9.0**	**17.8**	**48.3**	**12.7**	0.79	0.75	**38.5**	**43.6**	**50.0**	**45.1**

Table 1. Zero-shot captioning results on MSCOCO Karpathy-test split and NoCaps validations set. In, Near, and Out denote in-domain, near domain, and out-of-domain. MeaCap$_{TF}$ is the training-free version and MeaCap$_{ToT}$ is text-only training version.

Zero-shot captioning results on MSCOCO and NoCaps

Our Training-free version Compared with retrieve-based method

Methods	Text Corpus		MSCOCO						NoCap val (CIDEr)			
	Training	Memory	B@4	M	C	S	CLIP-S	BLIP2-S	In	Near	Out	Overall
ZeroCap [55]	✗	✗	2.6	11.5	14.6	5.5	0.87	0.70	13.3	14.9	19.7	16.6
Tewel *et al.* [54]	✗	✗	2.2	12.7	17.2	7.3	0.74	0.68	13.7	15.8	18.3	16.9
ConZIC [65]	✗	✗	1.3	11.2	13.3	5.0	**1.00**	0.76	15.4	16.0	20.3	17.5
CLIPRe [30]	✗	CC3M	4.6	13.3	25.6	9.2	0.84	0.70	23.3	26.8	36.5	28.2
DeCap [30]	CC3M	CC3M	8.8	16.0	42.1	10.9	0.76	-	34.8	37.7	49.9	39.7
MeaCap$_{TF}$	✗	CC3M	7.1	16.6	42.5	11.8	0.84	**0.81**	35.3	39.0	45.1	40.2
MeaCap$_{ToT}$	CC3M	CC3M	**9.0**	**17.8**	**48.3**	**12.7**	0.79	0.75	**38.5**	**43.6**	**50.0**	**45.1**

Table 1. Zero-shot captioning results on MSCOCO Karpathy-test split and NoCaps validations set. In, Near, and Out denote in-domain, near domain, and out-of-domain. MeaCap$_{TF}$ is the training-free version and MeaCap$_{ToT}$ is text-only training version.

Zero-shot captioning results on MSCOCO and NoCaps

Our Text-only-training version compared to previous text-only-training method

Methods	Text Corpus		MSCOCO						NoCap val (CIDEr)			
	Training	Memory	B@4	M	C	S	CLIP-S	BLIP2-S	In	Near	Out	Overall
ZeroCap [55]	✗	✗	2.6	11.5	14.6	5.5	0.87	0.70	13.3	14.9	19.7	16.6
Tewel *et al.* [54]	✗	✗	2.2	12.7	17.2	7.3	0.74	0.68	13.7	15.8	18.3	16.9
ConZIC [65]	✗	✗	1.3	11.2	13.3	5.0	**1.00**	0.76	15.4	16.0	20.3	17.5
CLIPRe [30]	✗	CC3M	4.6	13.3	25.6	9.2	0.84	0.70	23.3	26.8	36.5	28.2
DeCap [30]	CC3M	CC3M	8.8	16.0	42.1	10.9	0.76	-	34.8	37.7	49.9	39.7
MeaCap$_{TF}$	✗	CC3M	7.1	16.6	42.5	11.8	0.84	**0.81**	35.3	39.0	45.1	40.2
MeaCap$_{ToT}$	CC3M	CC3M	**9.0**	**17.8**	**48.3**	**12.7**	0.79	0.75	**38.5**	**43.6**	**50.0**	**45.1**

Table 1. Zero-shot captioning results on MSCOCO Karpathy-test split and NoCaps validations set. In, Near, and Out denote in-domain, near domain, and out-of-domain. MeaCap$_{TF}$ is the training-free version and MeaCap$_{ToT}$ is text-only training version.

In-domain / Cross domain captioning on MSCOCO and Flickr30K

In domain captioning

Methods	MSCOCO				Flickr30K			
	B@4	M	C	S	B@4	M	C	S
	Training on image-text pairs							
Bottom-Up [3]	36.2	27.0	113.5	20.3	27.3	21.7	56.6	16.0
OSCAR [31]	36.5	30.3	123.7	23.1	-	-	-	-
VinVL [66]	40.9	30.9	140.6	25.1	-	-	-	-
ClipCap [39]	33.5	27.5	113.1	21.1	-	-	-	-
SmallCap [47]	37.0	27.9	119.7	21.3	-	-	-	-
I-Tuning [37]	34.8	28.3	116.7	21.8	25.2	22.8	61.5	16.9
	Text-only-training, zero-shot inference							
ZeroCap† [55]	7.0	15.4	49.3	9.2	5.4	11.8	16.8	6.2
MAGIC [52]	12.9	17.4	49.3	11.3	6.4	13.1	20.4	7.1
ZEROGEN [56]	15.5	18.7	55.4	12.1	13.1	15.2	26.4	8.3
CLIPRe [30]	12.4	20.4	53.4	14.8	9.8	18.2	31.7	12.0
MeaCap$_{TF}$	9.1	20.6	56.9	15.5	7.2	17.8	36.5	13.1
MeaCap$_{ToT}$	**17.7**	**24.3**	**84.8**	**18.7**	**15.3**	**20.6**	**50.2**	**14.5**

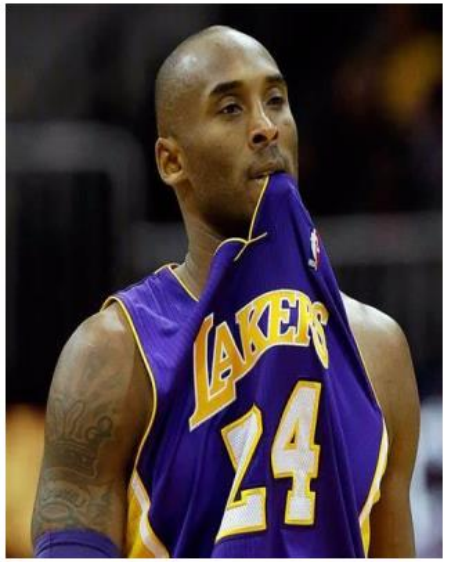
Table 2. In-domain captioning results on MSCOCO Karpathy-test split and Flickr30K Karpathy-test split. † means text-only re-implemented version from [52].

Cross domain captioning

Methods	MSCOCO → Flickr30k				Flickr30k→MSCOCO			
	B@4	M	C	S	B@4	M	C	S
MAGIC [52]	6.2	12.2	17.5	5.9	5.2	12.5	18.3	5.7
CLIPRe [30]	9.8	16.7	30.1	10.3	6.0	16.0	26.5	10.2
MeaCap$_{TF}$	7.1	16.6	34.4	11.4	7.4	16.2	46.4	11.2
MeaCap$_{ToT}$	**13.4**	**18.5**	**40.3**	**12.1**	**9.8**	**17.4**	**51.7**	**12.0**

Table 3. Cross domain captioning results on MSCOCO and Flickr30K Karpathy-test split.

Qualitative results with images contains world knowledge



ConZIC: A lakers player peeking through his sleeves prior to retiring.
ZeroCap: A great NBA star dead.
MAGIC:A man in a wetsuit with a wetsuit with a big.
DeCap: A man that is in the middle of a game with a tennis racket .

Memory Concepts

[basketball, shooting guard]

MeaCap$_{TF}$: The basketball star shooting guard Kobe.
MeaCap$_{ToT}$: The basketball star of shooting guard.

ConZIC: Triangular tower picture at virtual domain about france website.
ZeroCap: A French landmark is the name of the song.
MAGIC: A view of a big tower with a clock on it.
DeCap: A tower that is in the center of a tall tower .

[tower]

MeaCap$_{TF}$: The famous Eiffel tower in Paris.
MeaCap$_{ToT}$: The famous tower of tourist attraction.

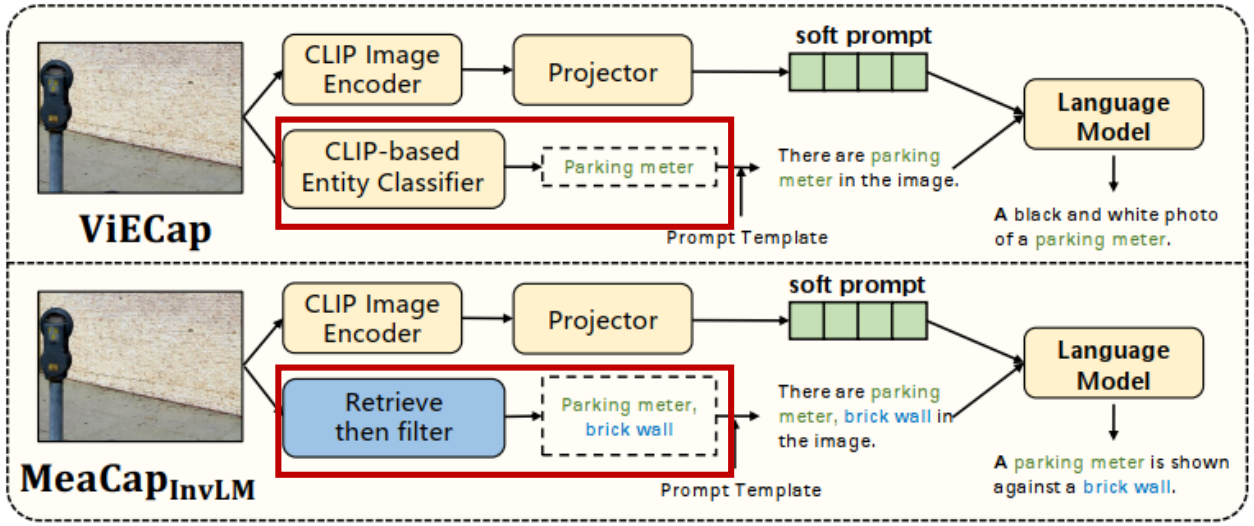
ConZIC: A dark knight representing a gray landscape background shaded.
ZeroCap: A Dark Knight in the film.
MAGIC: A black and white photo of a black and white zebra.
DeCap: A man that is standing in the dark.

[character, batman]

MeaCap$_{TF}$: A fictional character known as the batman.
MeaCap$_{ToT}$: A character of batman in the picture.

Apply designed memory mechanism to previous SOTA method

- Replace the entity classifier with our retrieve-then-filter module in a plug-and-play way



Methods	MSCOCO B@4	MSCOCO M	MSCOCO C	MSCOCO S	Flickr30K B@4	Flickr30K M	Flickr30K C	Flickr30K S
DeCap [30]	24.7	25.0	91.2	18.7	21.2	21.8	56.7	15.2
CapDec [41]	26.4	25.1	91.8	-	17.7	20.0	39.1	-
ViECap [13]	**27.2**	24.8	92.9	18.2	21.4	20.1	47.9	13.6
MeaCap$_{InvLM}$	**27.2**	**25.3**	**95.4**	**19.0**	**22.3**	**22.3**	**59.4**	**15.6**
	MSCOCO → Flickr30K				Flickr30K→MSCOCO			
DeCap [30]	16.3	17.9	35.7	11.1	12.1	18.0	44.4	10.9
CapDec [41]	17.3	18.6	35.7	-	9.2	16.3	27.3	-
ViECap [13]	17.4	18.0	38.4	11.2	12.6	19.3	54.2	12.5
MeaCap$_{InvLM}$	**18.5**	**19.5**	**43.9**	**12.8**	**13.1**	**19.7**	**56.4**	**13.2**

Table 4. In-domain and cross-domain captioning results with CLIP-invert language decoder.

Ablation for memory size

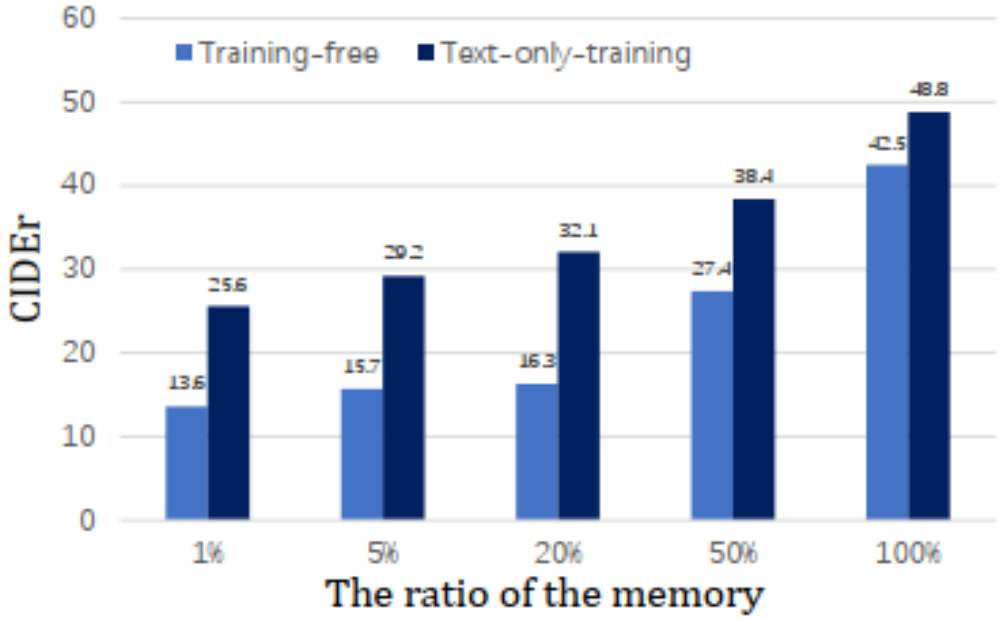


Figure 7. Ablation study on memory size.

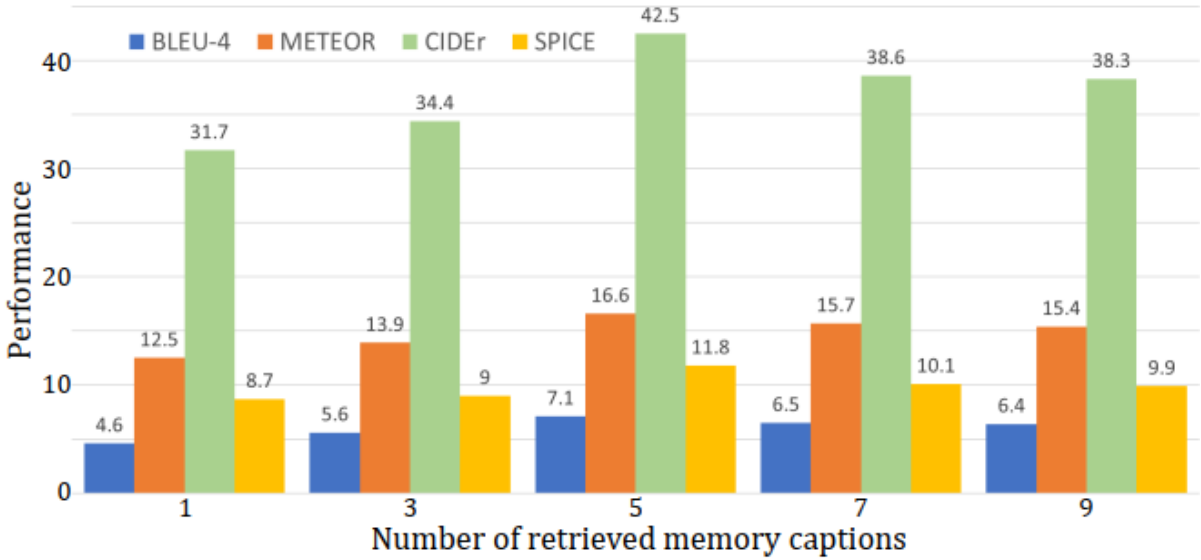


Figure 5. Effect of the number of retrieved memory captions. We reported the performance of $\text{MeaCap}_{\text{TF}}$ on the MSCOCO dataset with varying the number of retrieved memory captions.

Thanks for watching!