MeaCap: Memory-Augmented Zero-shot Image Captioning

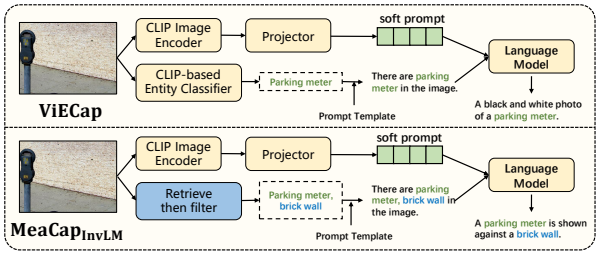


Figure 1. Difference between ViECap [2] and MeaCap$_{\text{InvLM}}$.

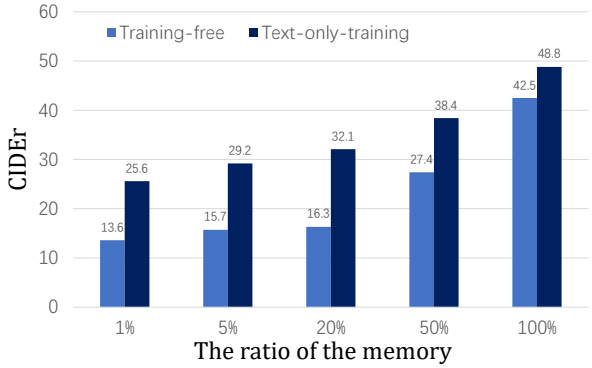


Figure 2. Ablation study on memory size.

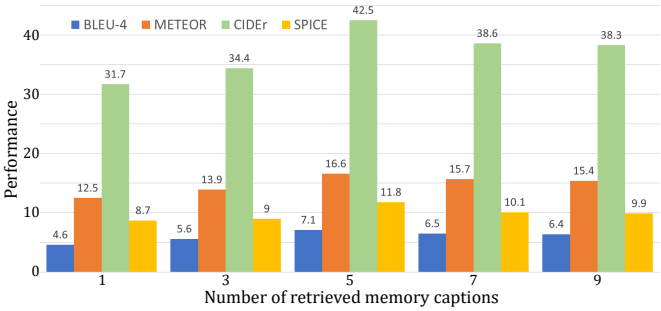


Figure 3. Effect of the number of retrieved memory captions. We reported the performance of MeaCap$_{\text{TF}}$ on the MSCOCO dataset with varying the number of retrieved memory captions.

A. More details of MeaCap with other LM

To demonstrate the effectiveness of our proposed memory concepts, we incorporate our proposed retrieve-then-filter module with the current SOTA text-only-training method ViECap [2], namely MeaCap$_{\text{InvLM}}$. ViECap is a text-only-training method that combines entity-aware hard prompts and learned visual soft prompts to generate image captions. The soft prompts are obtained by inputting the CLIP image embedding into the projector and the hard prompts are constructed by a CLIP-based entity classifier. As illustrated in Fig. 1, MeaCap$_{\text{InvLM}}$ remains the soft prompts branch and directly utilizes the pre-trained language model of ViECap. We only replace the entity classifier with our proposed retrieve-the-filter module to obtain key concepts and leverage the prompt template to inject the key concepts into a concept-aware sentence and get the hard prompts. Compared with the original ViECap which can only retrieve a single entity from pre-defined entity vocabulary, our proposed retrieve-then-filter module can extract more key visual concepts from the image, demonstrating that our proposed memory design can help to generate highly consistent descriptions with image content. Quantitative results under in-domain and cross-domain settings of MeaCap$_{\text{InvLM}}$ are shown in Tab. 4 in the main paper.

B. More analysis about the memory

In this section, we present some in-depth analysis about different aspects of external memory.

Numbers of retrieved memory captions. We also conduct experiments to investigate the impact of the number of retrieved memory captions as shown in Fig. 3. Our model achieves the best performance when retrieving five memory captions. As the number of retrieved memory captions exceeds five, there is a gradual performance degradation. The reason is all retrieved captions are considered equally regardless of the number of them. However, those captions with lower cosine similarity will bring more noise.

The size of external memory To explore the impact of the external memory size, we randomly sample different ratio text embeddings of CC3M as our external memory. The results are shown in Fig. 2. Overall, the training-free MeaCap$_{\text{TF}}$ and text-only-training MeaCap$_{\text{ToT}}$ both benefit from a large memory, and the performance increases as the memory size grows. It is due to that large memory usually covers more visual concepts and thus the retrieved captions are better aligned with the same image.

C. Zero-shot captioning with SS1M memory

To further explore the potential of MeaCap, we followed [5] and tried another textual memory SS1M. Partial comparison results are presented in Table 1, which illustrates the superior performance of MeaCap than DeCap.

Methods	Text Corpus		MSCOCO		NoCap val (CIDEr)			
	Training	Memory	M	C	In	Near	Out	Overall
DeCap	SS1M	SS1M	17.5	50.6	41.9	41.7	46.2	42.7
MeaCap$_{\mathrm{TF}}$	✗	SS1M	17.9	51.7	42.0	42.8	45.4	43.8
MeaCap$_{\mathrm{ToT}}$	SS1M	SS1M	**18.2**	**54.9**	**44.1**	**46.0**	**49.7**	**47.3**

Table 1. Zero-shot captioning results with SS1M memory.

D. BLIP2-S

BLIP-2 [4] is a large vision-language model that employs a frozen CLIP [6] image encoder and large language models [1, 7] to bootstrapping language-image pre-training on large image-text corpus. We use the pre-trained BLIP-2 and leverage the image encoder B_v and the text encoder B_t to compute cross-modality cosine similarity. Following CLIP-S [3], BLIP2-S is the product of a weight $w = 2$ and the image-text similarity, as:

$$\mathrm{BLIP2-S}(\mathbf{I}, \mathbf{T}) = w\mathrm{Cos}(B_v(\mathbf{I}), B_t(\mathbf{T})) \qquad (1)$$

where $\mathbf{I}$, and $\mathbf{T}$ are the image and the text. Cos denotes the cosine similarity.

E. Computational cost

Table 2 compares MeaCap with some SOTA methods w.r.t. the training and inference speed and the GPU memory usage. We can see that MeaCap shows competitive results with training-free methods and comparable ones with text-only methods.

Methods	Training-free			Text-only training		
	ZeroCap	ConZIC	**MeaCap**$_{\mathrm{TF}}$	DeCap	ViECap	**MeaCap**$_{\mathrm{ToT}}$
Training time (h)	0.0	0.0	0.0	2.1	8.1	11.0
Inference time (s)	140.2	9.1	1.4	0.9	0.7	1.4
Peak Memory (mb)	4121	2024	4063	2140	3407	4063

Table 2. Comparisons on computational costs and speed.

F. More samples and qualitatives.

Fig. 4 and Fig. 5 have shown some generated results with our proposed methods. For each image, we have shown the intermediate results of the retrieve-then-filter module, *i.e*. the retrieved memory captions, the extracted subject-predicate-object triplets, and the filtered key concepts. As we can see, we effectively filter the correct key concepts out. Based on the key concepts, MeaCap$_{\mathrm{TF}}$ and MeaCap$_{\mathrm{ToT}}$ can generate captions with high consistency with image content. Fig. 5 showcases that the extracted key concepts from memory contain much world knowledge and can also alleviate the knowledge-forgotten phenomenon of text-only-training methods.

WARNING: do not forget to delete the supplementary pages from your submission

References

[1] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[2] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3146, 2023. 1

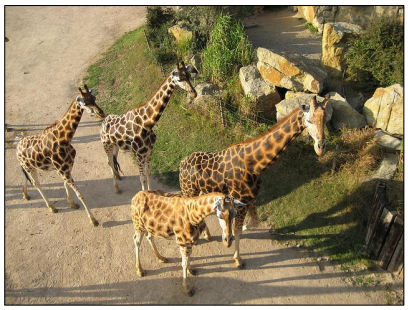
[3] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[5] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[7] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2



Retrieved memory captions:
1. a tower of giraffes with the male centered , all chewing.
2. group of giraffes in the zoo.
3. guests observe giraffes from a high platform.
4. group of giraffes at the zoo.
5. herd of giraffes in the zoo.

Subject-predicate-object triplets:
1. (male , is , chewing) , (giraffes , is , tower) , (male , center on top of , giraffes)
2. (giraffes , is , group of) , (giraffes , in , zoo)
3. (giraffes , on , platform) , (guests , observe , giraffes) , (platform , is , high)
4. (giraffes , is , group of) , (giraffes , at , zoo)
5. (giraffes , in , zoo)

Key concepts: ['giraffes', 'zoo']
MeaCap$_{\mathbf{TF}}$: The image depicts giraffes in a zoo enclosure surrounded.
MeaCap$_{\mathbf{ToT}}$: A group of giraffes in a zoo surrounded by stones.

Retrieved memory captions:
1. many people enjoy mushrooms on a pizza.
2. a pizza with black olives among the toppings.
3. pizza with mushrooms , tomato and olives in a box.
4. pizza with mushrooms , tomato and olives in a box.
5. close - up of vegetables on the pizza.

Subject-predicate-object triplets:
1. (people , enjoy , mushrooms) , (mushrooms , on , pizza)
2. (olives , is , black) , (olives , among , toppings) , (pizza , with , olives)
3. (pizza , in , box) , (pizza , with , tomatoes) , (pizza , with , olives) , (pizza , with , mushrooms)
4. (pizza , in , box) , (pizza , with , tomatoes) , (pizza , with , olives) , (pizza , with , mushrooms)
5. (vegetables , on , pizza)

Key concepts: ['vegetables', 'olives', 'toppings', 'box']
MeaCap$_{\mathbf{TF}}$: The image depicts a pizza made of vegetables, olives and toppings in a box.
MeaCap$_{\mathbf{ToT}}$: A pizza with vegetables, olives and other toppings in a box.

Retrieved memory captions:
1. children reading the book on picnic in summer park.
2. children reading the book on picnic in summer park.
3. with young kids , dinner can feel like feeding time at the zoo.
4. young friends eating pizza in the park.
5. young friends eating pizza in the park.

Subject-predicate-object triplets:
1. (children , on , picnic) , (children , in , park) , (park , is , summer) , (children , read , book)
2. (children , on , picnic) , (children , in , park) , (park , is , summer) , (children , read , book)
3. (kids , is , young) , (kids , at , zoo) , (dinner , at , zoo) , (dinner , have , kids)
4. (friends , eat , pizza) , (friends , is , young) , (friends , in , park)
5. (friends , eat , pizza) , (friends , is , young) , (friends , in , park)

Key concepts: ['children', 'dinner', 'park']
MeaCap$_{\mathbf{TF}}$: A picture showing two children eating dinner at a park.
MeaCap$_{\mathbf{ToT}}$: Young children eating dinner in the park.

Retrieved memory captions:
1. person smiling with a big plate of chocolate cake and vanilla ice cream in front of him.
2. portrait of a young man eating chocolate cake with a fork.
3. young man eating a slice of chocolate cake with a fork.
4. a man smiles , a chocolate cake on the table in front of him.
5. a man smiles , a chocolate cake on the table in front of him.

Subject-predicate-object triplets:
1. (plate , is , big) , (ice cream , is , vanilla) , (person , is , smiling) , (plate , in front of , person) , (cake , on , plate) , (cake , is , chocolate) , (person , is , smiling)',
2. (man , is , young) , (man , eat , chocolate cake) , (man , with , fork)
3. (man , is , young) , (man , eat , chocolate cake) , (chocolate cake , is , slice) , (man , eat with , fork)
4. (man , is , smiling) , (chocolate cake , on , table) , (table , in front of , man)
5. (man , is , smiling) , (chocolate cake , on , table) , (table , in front of , man)

Key concepts: ['young man', 'chocolate cake', 'table', 'cafe']
MeaCap$_{\mathbf{TF}}$: There are pictures of a young man smiling on holiday eating chocolate cake on the table in the cafe.
MeaCap$_{\mathbf{ToT}}$: A young man smiling and eating chocolate cake on a table in the cafe.

Figure 4. More qualitative results.

Retrieved memory captions:
1. everything you need to know about the luxury sedan.
2. v8 portfolio : super long , light and performing sedan at a supercharged price.
3. first drive : adds sedan to its line ς up.
4. new black luxury sedan turning slightly to the right.
5. the car is the clearest idea yet of what to expect from the next generation luxury sedan.
Subject-predicate-object triplets:
1. (sedan , is , luxury)
2. (sedan , is , long) , (sedan , is , performing) , (sedan , is , supercharged) , (sedan , is , sedan)
3. (sedan , in , line)
4. (sedan , is , black) , (sedan , is , luxury) , (sedan , is , new) , (sedan , is , turning)
5. (sedan , is , luxury) , (sedan , is , next generation) , (car , is , clearest) , (car , is , idea)
Key concepts: ѓ'ÌÌÕØÕÒÙ ÔÅÄÁÎэ'
'Ë'ĦĦ'ÄĦ'Ì$_{ñ\acute{e}}$: The image depicts a luxury sedan with the BMW logo.
'Ë'ĦĦ'ÄĦ'Ì$_{ñ\check{r}ñ}$: The BMW luxury sedan shown here.

Retrieved memory captions:
1. iron man , pictured here , was one of the more overhauled heroes , due to severely outdated design.
2. actor takes the stage in a new poster for iron man.
3. iron man has a new poster with actor in some heat in film.
4. costume designer created a new suit for the film.
5. what do you think of the new poster for iron man?
Subject-predicate-object triplets:
ρ¢ ĵ Í ÁÎ ĥÉÓĥÉÒÏ Î Qĥĵ ÄÃÓÉÇÎ ĥÉÓĥÔÅÕÄÔÅÀÌÙ ÏÕÔÄÁÔÅÄ Qĥĵ Í ÁÎ ĥÈÁÖÃĥÄÄÓÉÇÎ Qĵĥ
ς¢ ĵ ÐÏÓÔÅÔĥÉÓĥ ÎÅ× Qĥĵ ÐÏÓÔÅÔĥÉÓĥÉÒÏÎ Í ÁÎ Qĥĵ ÁÃÔÏÔĥÔÁÉÊÃĥÓÔÁÇÅ Qĥĵ ÃÃÔÏÔĥÉÎ ĥÐÏÓÔÅÔ Qĵĥ
σ¢ ĵ ÐÏÓÔÅÔĥÉÓĥ ÎÅ× Qĥĵ Í ÁÎ ĥÉÓĥÉÒÏÎ Qĥĵ Í ÁÎ ĥÈÁÖÅĥÐÏÓÔÅÔ Qĥĵ Í ÁÎ ĥÉÎ ĥÆÉÌ Í Qĥĵ Í ÁÎ ĥÉÎ ĥÈÃÁÔ Qĵĥ
τ¢ ĵ ÓÕÉÔĥÉÓĥ ÎÅ× Qĥĵ ÃÏÓÔÕ Í Å ÄÃÓÉÇÎ ÅÒĥ ÃÒÃÁÔÅĥ ÓÕÉÔ Qĥĵ ÓÕÉÔĥÉÎ ĥÆÉÌ Í Qĵĥ
υ¢ ĵ ÐÏÓÔÅÔĥÉÓĥ ÎÅ× Qĥĵ ÐÏÓÔÅÔĥÉÓĥÉÒÏÎ Í ÁÎ Qĵ
Key concepts: ѓ'Î Å× ÐÏÓÔÅÔĵĵ ÍÉÒÏÎ Í ÁÎэ'
'Ë'ĦĦ'ÄĦ'Ì$_{ñ\acute{e}}$: The image depicts the new poster as an iron man acting like a hero.
'Ë'ĦĦ'ÄĦ'Ì$_{ñ\check{r}ñ}$: The new movie poster for iron man.

Retrieved memory captions:
1. replica of the statue of liberty wrapped in flag stock photo ς 20491789.
2. did you know the statue of liberty was once a dull copper color.
3. the statue of liberty contains a large amount of copper.
4. visit the statue of liberty during your itinerary.
5. visit the statue of liberty during your itinerary.
Subject-predicate-object triplets:
1. (statue , wrap in , flag) , (flag , is , stock photo) , (statue , is , replica)
2. (statue , is , copper) , (statue , is , dull)
3. (statue , contain , copper)
4. (statue of liberty)
5. (statue of liberty)
Key concepts: ['statue of liberty']
'Ë'ĦĦ'ÄĦ'Ì$_{ñ\acute{e}}$: A photo of the statue of liberty in America taken at sunset.
'Ë'ĦĦ'ÄĦ'Ì$_{ñ\check{r}ñ}$: The staue of liberty at sunset.

Figure 5. More qualitative results with world knowledge.