Supplementary Material for EVCap: Retrieval-Augmented Image Captioning with External Visual–Name Memory for Open-World Comprehension

This supplementary material complements our paper with the following sections: First, we delve into the implementation specifics of our EVCap, which were not covered in the main paper (see Sec. A). Second, we offer an expanded discussion on the external visual-name memory, as utilized in the main paper (see Sec. B). Finally, we present additional results to evaluate the effectiveness of EVCap (see Sec. C).

A. Implementation Details

Customized Q-Former. Fig. A depicts customized Q-Former, where the image embedding port (CLIP visual feature port) now receives retrieved object names $\mathcal{S}$, while the text port receives visual features $\mathcal{Q}$.

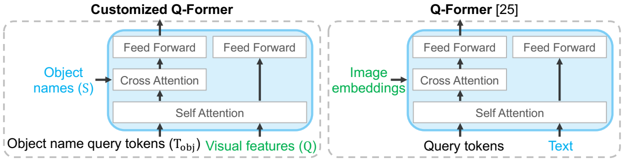


Figure A. Customized Q-Former and original Q-Former [25].

Implementation. Our method is based on Pytorch and is trained within one epoch with a batch size of 24 using mixed precisions. We optimize the model using AdamW, setting the weight decay at 0.05, and using β_1 and β_2 values of 0.9 and 0.99, respectively. A cosine learning rate (LR) decay strategy is adopted, starting with an initial LR of 1e-4. The model undergoes 5000 linear warm-up steps, beginning with a start LR of 1e-6. During the evaluation phase, we use a beam search strategy with a beam size of 5 to generate captions.

B. External Visual–Name Memory

B.1. LVIS memory

As stated in Sec. 3.2 of the main paper, we utilize 1203 objects from the LVIS dataset. For each of these objects, we randomly select between one and ten images from LVIS. Additionally, we enrich our data by incorporating five synthetic images for each object, created using stable diffusion. We show two samples of this external visual-name memory, constructed using objects from LVIS in Fig. B.

B.2. WHOOPS memory

To illustrate the scalability of the external memory in EVCap, we expand it by integrating WHOOPS knowledge into the original external visual–name memory in Sec. 5.2 and Sec. 5.3 of the main paper. Specifically, we focus on objects that are mentioned in the answers of VQA annotations in the WHOOPS dataset because of their conciseness and emphasis on key objects. For each of these objects, we produce five synthetic images employing stable diffusion. Two examples from this augmented memory, featuring newly added object images and their corresponding names, are presented in Fig. C.

Table A. Quantitative results under the CIDEr score on in-domain (In), near-domain (Near), out-domain (Out), and overall data of the NoCaps test set. * denotes using a **memory bank**. We note that our results on the test set are publicly submitted to Nocaps leader-board[a] (8th rank). Higher score is better. **Bold** indicates the best results among compared methods, normal indicates the second best results.

Method	In	Near	Out	Overall
Heavyweight-training models				
VinVL [45]	93.8	89.0	66.1	85.5
NOC-REK* [40]	100.0	95.7	77.4	93.0
OSCAR [26]	81.3	79.6	73.6	78.8
Lightweight-training models				
SmallCap* $_{\text{GPT2}}$ [35]	87.9	84.6	84.4	85.0
EVCap* $_{\text{Vicuna-13B}}$	**114.9**	**117.0**	**117.1**	**116.8**
Specialist SOTAs				
CogVLM $_{\text{Vicuna-7B}}$ [5]	–	–	128.0	126.4
PaLI $_{\text{mT5-XXL}}$ [9]	–	–	–	124.4
PaLI-X $_{\text{UL2-32B}}$ [8]	–	–	–	124.3

[a] https://eval.ai/web/challenges/challenge-page/355/leaderboard/1011

C. Additional Results

C.1. Experiments on NoCaps test set

We additionally assess our EVCap against SOTAs on the NoCaps test set, since we notice several other methods have also benchmarked their performance on this dataset. Note that, NoCaps test set does not have publicly accessible ground truth annotations. To obtain evaluation scores, we submitted our results to the NoCaps leaderboard.

Quantitative results. Tab. A presents the quantitative results of our EVCap on the NoCaps test set. Our method outperforms all other SOTA models, both in heavyweight-training and lightweight-training categories, that have re-

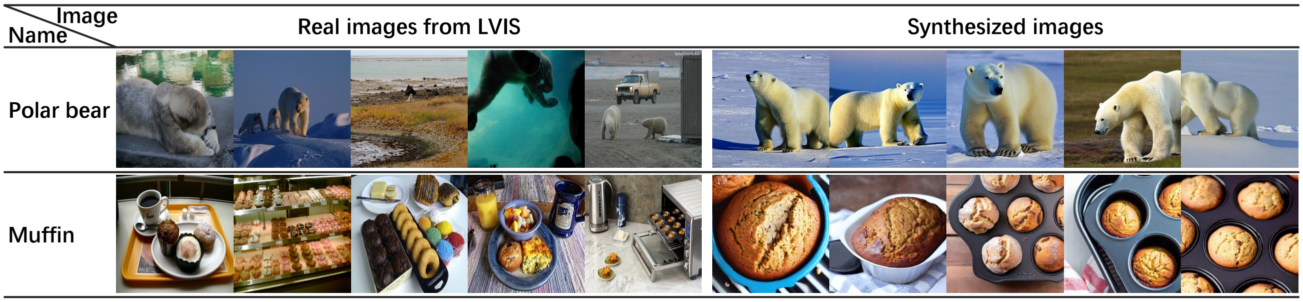


Figure B. Samples of images and corresponding names in the external visual–name memory constructed from LVIS's objects.

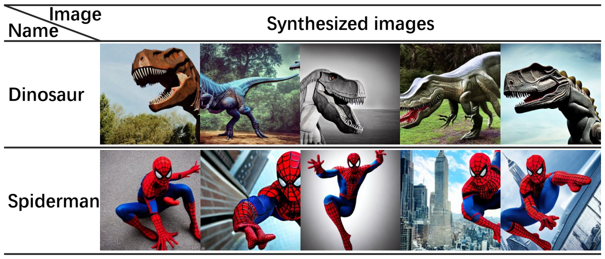


Figure C. Samples of images and corresponding names in the added WHOOPS's objects.

ported results on this dataset. Additionally, as a lightweight method, our approach achieves the 8th rank on the NoCaps leaderboard, only surpassed by specialized SOTAs such as CogVLM, which holds the 1st rank.

Qualitative results. Fig. E shows the captions generated by our EVCap alongside those from three SOTA methods on the NoCaps test set. It also includes the object names retrieved by EVCap and the captions retrieved by SmallCap. Consistent with the findings in the main paper, SmallCap tends to generate hallucinatory objects that are absent in the input images, such as *"tie"* and *"mouse"*. The same hallucinatory object *"mouse"* is also found in the retrieved captions, indicating that SmallCap's diminished performance is largely due to its reliance on retrieved captions containing irrelevant information. In comparison, our EVCap demonstrates a performance on par with BLIP-2.

C.2. Further analysis

Comparison on training time and used GPUs. Tab. B compares training time and the used GPUs of our EVCap with various SOTA models. Due to the diversity of GPUs employed across different models, drawing a direct comparison is challenging. Nevertheless, it's evident that the training time for our EVCap is comparatively shorter than most models.

Number of object name query tokens. We explore the impact of varying the number of retrieved object names P (Sec. 3.4 of the main paper) on EVCap in Fig. D. We train

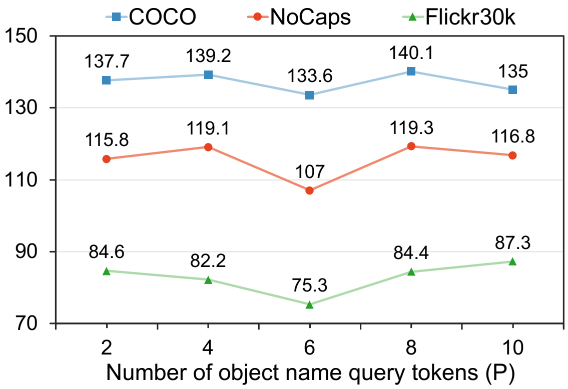


Figure D. CIDEr scores after training EVCap with the number of object name query tokens P from 2 to 10. The results indicate that the performance is relatively optimal when P is set to be 8.

Table B. Comparison against SOTA methods on training time and used GPUs. * denotes using a **memory bank**.

Method	Training time	GPUs
Heavyweight-training models		
VinVL [45]	–	–
AoANet+MA* [16]	–	–
NOC-REK* [40]	8d	2 RTX3090
RCA-NOC*[13]	1d	8 A100
ViECap $_{\text{GPT2}}$[15]	–	–
InstructBLIP $_{\text{Vicuna-13B}}$[11]	1.5d	16 A100
OSCAR [26]	74h	1 V100
BLIP [24]	–	2 16-GPU nodes
BLIP-2 $_{\text{FlanT5-XL}}$ [25]	~9d	16 A100
REVEAL* $_{\text{T5}}$ [20]	5d	256 CloudTPUv4 chips
Lightweight-training models		
MiniGPT4 $_{\text{Vicuna-13B}}$ [46]	10h	4 A100
SmallCap* $_{\text{GPT2}}$ [35]	8h	1 A100
ClipCap $_{\text{GPT2}}$ [29]	6h	1 GTX1080
EVCap* $_{\text{Vicuna-13B}}$	3h	4 A6000
Specialist SOTAs		
Qwen-VL $_{\text{Qwen-7B}}$ [5]	–	–
CogVLM $_{\text{Vicuna-7B}}$ [41]	1d	4096 A100
PaLI $_{\text{mT5-XXL}}$ [9]	–	–
PaLI-X $_{\text{UL2-32B}}$ [8]	–	–

SmallCap: A group of snowboarders standing in the snow.
MiniGPT4: A group of people standing in the snow wearing ski gear and smiling.
BLIP-2: A group of people standing in the snow on a ski lift.
EVCap: Two skiers are posing for a picture in the snow.
(glove, ski pole, jacket, snowboard, helmet, ski, map, goggles, ski boot, backpack)

SmallCap' retrieved captions:
Two men with snowboarding gear standing on slope next to a mountain.
A couple of snowboarders that are in the snow.
A pair of men standing on snowboards giving the thumbs up.
A snowboarder in a green jacket and a skier.

SmallCap: A man in a suit and tie speaking at a podium.
MiniGPT4: A man standing at a podium in front of a crowd.
BLIP-2: A man giving a speech at a podium in front of a crowd.
EVCap: A man standing behind a podium giving a speech.
(dress, identity card, television set, ring, spectacles, handkerchief, cymbal, tablecloth, fire alarm, plastic bag)

SmallCap' retrieved captions:
A large man speaking in front of a crowd.
A man at a podium speaking at an event.
A person speaking something with a mike in his hands.
A man is speaking while a picture is taken of him.

SmallCap: A computer mouse sitting next to a keyboard.
MiniGPT4: A red joystick sitting on top of a keyboard.
BLIP-2: A computer mouse on a desk next to a keyboard.
EVCap: A red joystick sitting next to a computer keyboard.
(mouse computer equipment, coaster, place mat, joystick, router computer equipment, combination lock, money, battery, notebook, cat)

SmallCap' retrieved captions:
An orange and white electric gadget sits on a red surface.
A white joystick held by a child for a video game.
A video game control is held by grips.
An electronic gadget connected to a keyboard and wireless mouse.

Figure E. Examples of captions generated by our EVCap, and three SOTAs on the NoCaps test set. We also list the retrieved object names by our EVCap (below the captions of EVCap) and retrieved captions by SmallCap (right side) in *italics*.

the model using different values of P, ranging from 2 to 10, and evaluate the performance under CIDEr on all three benchmarks. The results suggest that setting $P = 8$ offers relatively optimal results.

C.3. More qualitative examples.

More qualitative examples on the COCO test set, NoCaps validation set, and Flickr30k test set are shown in Fig. F, Fig. G, and Fig. H, respectively.

GT: Some sheep that are grazing in a snowy field.
SmallCap: A herd of sheep standing in the snow.
MiniGPT4: A herd of sheep grazing in a snow covered field.
BLIP-2: A herd of sheep grazing in a field near a house.
EVCap: A herd of sheep grazing in a snow covered field.
(lamb animal, goat, deer, bull, ram animal, calf, hammock, cow, bear, giraffe)

SmallCap’ retrieved captions:
A flock of sheep are gathered at the corner of a cabin in a field.
A group of sheep standing in a snowy yard outside a barn.
Sheep graze in a field next to two large houses.
A farm setting with sheep grazing in the yard area.

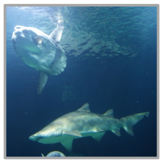
GT: Two rectangular slices of pizza with multiple toppings.
SmallCap: Two slices of pizza sitting on top of a white plate.
MiniGPT4: Two slices of pizza on a metal tray with cheese and vegetables.
BLIP-2: Two slices of pizza on a white plate.
EVCap: Two slices of pizza topped with cheese and vegetables.
(mushroom, salami, crouton, sausage, quesadilla, bell pepper, lettuce, squid food, crescent roll, avocado)

SmallCap’ retrieved captions:
A small slice of gourmet flat bread pizza.
Flat bread pizza slices piled on top of each other.
A little personal sized pizza cut into squares.
A collection of differently topped pizza slices on a plate.

GT: There are crafts and craft supplies on a table.
SmallCap: A number of craft items sitting on a table.
MiniGPT4: A pair of scissors and some yarn on a table.
BLIP-2: A blue pair of scissors sits on a table next to other items.
EVCap: A pair of scissors sitting next to a crocheted item.
(mushroom, salami, crouton, sausage, quesadilla, bell pepper, lettuce, squid food, crescent roll, avocado)

SmallCap’ retrieved captions:
Two cellphones have cute homemade cellphone covers.
Several crocheted items with yarn scissors and crochet hooks.
A series of crafting tools are laid out mostly for sewing.
A number of craft items sitting on a table.

Figure F. Examples of captions generated by our EVCap, and three SOTAs on the COCO test set. We also list the retrieved object names by our EVCap (below the captions of EVCap) and retrieved captions by SmallCap (right side) in *italics*.



GT: A shark is swimming in dark blue water near other fish.
SmallCap: A couple of fish that are in the water.
MiniGPT4: A large shark swimming in the ocean with other fish.
BLIP-2: A shark and a turtle swimming in the water.
EVCap: A shark and a fish swim in the water.
(polar bear, manatee, rhinoceros, cow, grizzly, crossbar, foal, shark, ostrich, crow)

SmallCap’ retrieved captions:
Two large animals standing in a pool of water.
A person looks into an aquarium at a large animal swimming.
This is a large fish that is in an aquarium.
Some animals that are in the water together.

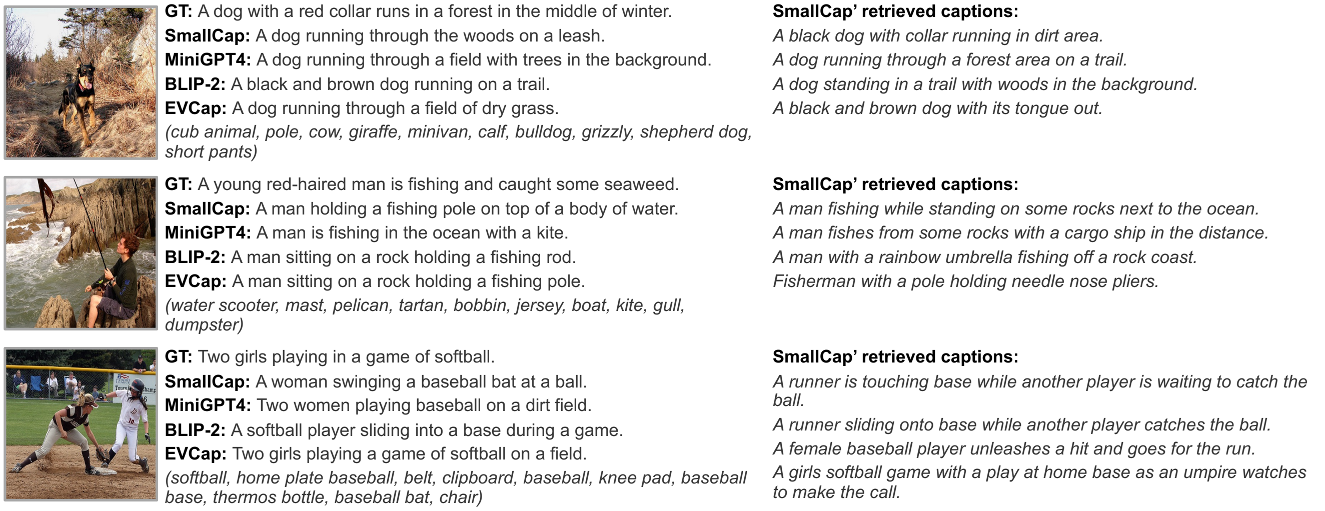
GT: A sink, a tub with towels, soap and other bath tub accessories.
SmallCap: A bathroom with a sink and a toilet.
MiniGPT4: A bathroom with a sink and a shower in it.
BLIP-2: A white bathroom with a sink and a shower.
EVCap: A bathroom with a sink, shower and towels.
(sink, urinal, shower head, hose, washbasin, towel rack, radiator, cover, hair dryer, shower curtain)

SmallCap’ retrieved captions:
An image of a bathroom setting with double faucet.
A very nice looking clean cut and contemporary styled wash basin.
A modern bathroom with a toilet and pedestal sink.
A washroom area with a sink soap and glasses.

GT: A person holding a ipod and a iphone in both hand.
SmallCap: A person holding a cell phone in their hand.
MiniGPT4: A person holding an ipod in their hand.
BLIP-2: A person holding an ipod next to a cell phone.
EVCap: A person holding an iPhone next to an iPod.
(ipod, cellular telephone, sword, olive oil, boxing glove, toaster, stapler stapling machine, rat)

SmallCap’ retrieved captions:
An apple computer ipod and other electronic devices.
A variety of apple ipod products on display.
Multiple ipods on a desk surrounded by computers.
Man at computer holding ipod and interfacing the two devices.

Figure G. Examples of captions generated by our EVCap, and three SOTAs on the NoCaps validation set. We also list the retrieved object names by our EVCap (below the captions of EVCap) and retrieved captions by SmallCap (right side) in *italics*.



GT: A dog with a red collar runs in a forest in the middle of winter.
SmallCap: A dog running through the woods on a leash.
MiniGPT4: A dog running through a field with trees in the background.
BLIP-2: A black and brown dog running on a trail.
EVCap: A dog running through a field of dry grass.
(cub animal, pole, cow, giraffe, minivan, calf, bulldog, grizzly, shepherd dog, short pants)

SmallCap' retrieved captions:
A black dog with collar running in dirt area.
A dog running through a forest area on a trail.
A dog standing in a trail with woods in the background.
A black and brown dog with its tongue out.

GT: A young red-haired man is fishing and caught some seaweed.
SmallCap: A man holding a fishing pole on top of a body of water.
MiniGPT4: A man is fishing in the ocean with a kite.
BLIP-2: A man sitting on a rock holding a fishing rod.
EVCap: A man sitting on a rock holding a fishing pole.
(water scooter, mast, pelican, tartan, bobbin, jersey, boat, kite, gull, dumpster)

SmallCap' retrieved captions:
A man fishing while standing on some rocks next to the ocean.
A man fishes from some rocks with a cargo ship in the distance.
A man with a rainbow umbrella fishing off a rock coast.
Fisherman with a pole holding needle nose pliers.

GT: Two girls playing in a game of softball.
SmallCap: A woman swinging a baseball bat at a ball.
MiniGPT4: Two women playing baseball on a dirt field.
BLIP-2: A softball player sliding into a base during a game.
EVCap: Two girls playing a game of softball on a field.
(softball, home plate baseball, belt, clipboard, baseball, knee pad, baseball base, thermos bottle, baseball bat, chair)

SmallCap' retrieved captions:
A runner is touching base while another player is waiting to catch the ball.
A runner sliding onto base while another player catches the ball.
A female baseball player unleashes a hit and goes for the run.
A girls softball game with a play at home base as an umpire watches to make the call.

Figure H. Examples of captions generated by our EVCap, and three SOTAs on the Flickr30k test set. We also list the retrieved object names by our EVCap (below the captions of EVCap) and retrieved captions by SmallCap (right side) in *italics*.