A. Cross-Attention Layers

We add cross-attention layers to GPT-2 following Vaswani *et al.* [37]. A set of queries Q, values V and keys K are processed by multi-head cross-attention (MHA) with h heads as follows:

$$\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_i, ..., \text{head}_h)\mathbf{W}_O \tag{2}$$

$$\text{head}_i = \text{Att}(\mathbf{QW}_i^Q, \mathbf{KW}_i^K, \mathbf{VW}_i^V) \tag{3}$$

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\frac{\mathbf{QK}^T}{\sqrt{d_k}}\mathbf{V}, \tag{4}$$

where $\mathbf{W}_i^K \in \mathbb{R}^{d_{encoder} \times d}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_{encoder} \times d}$, $\mathbf{W}_i^Q \in \mathbb{R}^{d_{decoder} \times d}$, and $\mathbf{W}_O \in \mathbb{R}^{h*d \times d_{decoder}}$ are learned model parameters, and the attention dimensionality d is set manually to a desired value.

We explore different values for the dimensionality of the cross-attention projection matrices (d) to achieve a lower number of trainable parameters, as discussed in Section 6.3.

B. Design Choices and Hyperparameters

Retrieval encoder	CIDEr
ViT-B/32	109.5
ViT-L/14	115.2
ResNet-50x4	114.1
ResNet-50x64	117.9

(a) CLIP version for retrieval

Image encoder	CIDEr
ViT-B/32	117.9
ResNet-50x64	107.5

(b) CLIP version in main model

k	CIDEr
1	113.38
2	116.03
3	117.47
4	117.88
5	117.87
6	117.82

(c) k for retrieval

Table 7. Hyperparameter tuning of the retrieval mechanism, measuring CIDEr on the validation set of COCO.

We developed the optimal configuration for SmallCap by first tuning the retrieval encoder, then the main vision encoder, followed by the number of retrieved captions, and lastly the cross-attention dimensionality. The results from the first three steps are presented below, while the last step is presented in Section 6.3. At the start of the tuning process, the main vision encoder was set to CLIP-ViT-B/32, the number of retrieved captions to 5 and the cross-attention dimensionality to 64.

B.1. Retrieval Encoder

We compared three CLIP versions for retrieval. As seen in Table 7 (a), CLIP-ResNet-50x64 performs best so we used this encoder for the final SmallCap model.

B.2. Main Vision Encoder

Next, we compared the use of CLIP-ResNet-50x64 to CLIP-ViT-B/32 as vision encoder in the main model. To use CLIP-ResNet-50x64 as an image encoder, we had to add a linear projection to match the dimensionality of the encoder to that of the decoder for the purposes of cross-attention. In Table 7 (b), we observe that CLIP-ViT-B/32 has better performance.

B.3. Number of Retrieved Captions

We also tuned the number of retrieved captions, training the model with k ranging from 1 to 6. Results are reported in Table 7 (c) and indicate that $k = 4$ is the optimal value. Qualitative analysis also showed that it is important to retrieve a sufficient number of captions since retrieving more captions can make the model more robust against wrong information from certain retrieved captions, as depicted in the second example in Figure 3.

C. Prompt

Besides the template proposed in Section 3.2, we explored other templates for prompting, including different separators between the retrieved captions (e.g., comma, dot, empty lines). However, we found that the prompt template has little impact on the model's performance, in line with previous work [13]. The final template we used was:

```
Similar images show\n\n<caption
1>\n\n<caption 2>\n\n<caption
3>\n\n<caption 4>.\n\nThis image shows
```

D. `nocaps`

In Table 8, we show results on the `nocaps` validation set, since several recent studies only include performance on the validation set, following [19]. In line with the test set results, SmallCap$_{+W+H}$ outperforms other lightweight-training models and even the large model OSCAR, especially in the *Out*-of-domain setting.

E. SmallCap with Alternative Decoders

E.1. Larger GPT-2 decoders

In Figure 7, we study the scaling behaviour of SmallCap with larger decoders for different cross-attention dimensionalities ($d = 4$, $d = 8$ and $d = 16$). We can see

Models	In	Near	Out	Entire
		Validation		
OSCAR$_{\text{Large}}$*	78.8	78.9	77.4	78.6
I-Tuning$_{\text{Large}}$$^{\circ}$	89.6	80.4	64.8	78.5
I-Tuning$_{\text{Medium}}$$^{\circ}$	89.6	77.4	58.8	75.4
ClipCap*	84.9	66.8	49.1	65.8
SMALLCAP	87.6	78.6	68.9	77.9
SMALLCAP$_{\text{+W+H}}$	**90.5**	**85.6**	**91.5**	**87.5**

Table 8. Validation results in CIDEr score on `nocaps`. * Results copied from the respective publications. ⋆ Results computed by us. ○ Results obtained through personal communication.

	OPT-125M	OPT-350M
With retrieval	120.8	120.8
Without retrieval	113.4	112.6

Table 9. Validation results in CIDEr score on COCO.

it is beneficial to train with GPT-Medium and GPT-Large across the different dimensionalities of the cross-attention. Controlling the cross-attention dimensionality allows us to leverage these larger decoders without a massive increase in the number of trainable parameters while maintaining a stable performance. Notwithstanding, a larger decoder still requires more GPU memory, which means that we had to reduce the batch size and use gradient accumulation to train SMALLCAP $_{\text{Medium}}$ and SMALLCAP $_{\text{Large}}$ models.

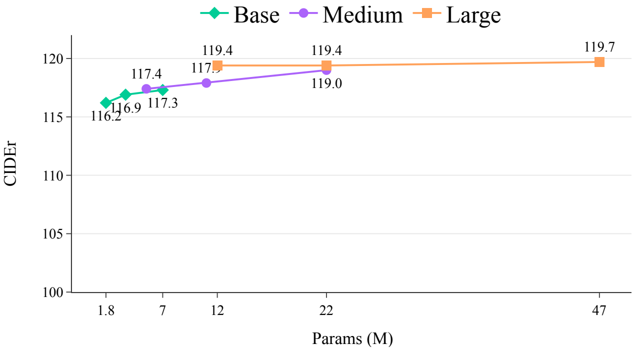


Figure 7. CIDEr performance on the COCO validation set across different decoder sizes: GPT-Base, GPT-Medium, GPT-Large and cross-attention dimensionalities $d = 4, 8, 16$.

E.2. OPT decoders

In Section 6.4, we also showed results with SMALLCAP variants trained with a different decoder based on OPT [48]. In Table 9 we report results from models trained with and without retrieval. The large drop in performance without retrieval demonstrates that retrieval is key to the good model performance here too, as was observed for SMALLCAP using a GPT-2 decoder (see Section 6.3).

Dataset	Data type	Size
Web [18]	Image captions	12M
Human-Labeled		2.1M
COCO [7]	Image captions	566K
Flickr [47]	Image captions	145K
VizWiz [10]	Image captions	117K
LN Ade20k [28]	Image Narratives	19K
LN COCO [28]	Image Narratives	121K
LN Flick30k [28]	Image Narratives	28K
LN Open Images [28]	Image Narratives	496K
MSR-VTT [45]	Video captions	130K
VATEX [41]	Video captions	349K
TGIF [20]	GIF captions	125K
Clotho [9]	Audio captions	14K

Table 10. Data used in the datastore for the experiments reported in Section 3.2 along with size in terms of image-caption pairs. LN stands for Localized Narratives [28].

F. Data

For the experiments in Section 5, we explored different sources of data to include in the datastore, detailed in Table 10. Specifically, we used the cleaner web data version proposed in Li *et al.* [18], which contains synthetic model-generated texts for the same web images, instead of using the original noisy web texts given the findings that noisy web texts are suboptimal for vision-and-language tasks [18]. We also used different human-labeled data beyond image captioning datasets, including video captioning, audio captioning and localized narratives. We only included in the datastore text with length shorter than 25 tokens. Regarding the index, for human-labeled data, since it is limited-scale, we used `IndexFlatIP` without requiring training. For the web data, given its larger size, we used `IndexIVFFlat` with a training stage to speed up the search (with the hyperparameter *nprobe* equal to 16). In terms of space, the COCO datastore takes up 2.2GB, the Human-Labeled datastore takes 8GB, and the Web datastore takes 49GB. Future work can include a further exploration of index types, since the FAISS library provides different indexes to customize for a faster search and lower memory footprint (e.g., through quantization).

G. Inference time

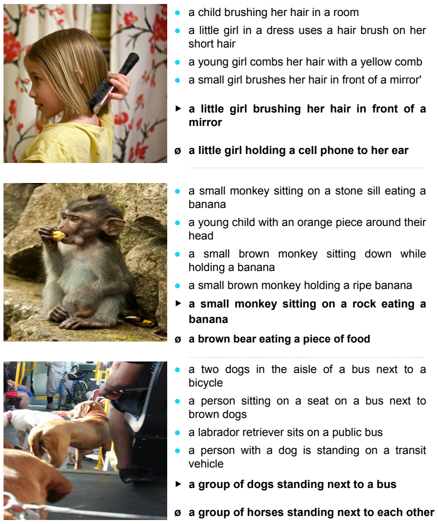
SMALLCAP is a lightweight-training captioning model. Although training efficiency is of crucial importance, especially in contexts involving limited resources, inference time should also be to taken into account. We thus measured the inference time of SMALLCAP and CaMEL on an NVIDIA A100 GPU across 1,000 randomly sampled images from COCO. The resulting values are 0.22 and 0.58 seconds per image, respectively, i.e., SMALLCAP is much faster than CaMEL, likely due to CaMEL's dual decoder architecture. In Section 4.1, we also report the residual difference of generating a caption with and without retrieval at inference time.

H. More Qualitative Examples

Figure 8 shows examples of captions generated by SMALLCAP on the COCO dataset, compared to a variant trained without retrieval. In line with the quantitative results that were presented before, SMALLCAP can better describe an input image when conditioning on the retrieved examples. In the first picture, we see that without retrieval a brush is mistaken for a cell phone, which is a more common object in the COCO training data.

In addition, in Figure 9, we provide more examples of how SMALLCAP adapts to Flickr30k, VizWiz, and MSR-VTT, by replacing the contents of the datastore with the in-domain data.

Lastly, we measured the importance of generating a caption conditioned on retrieved information compared to directly using the nearest caption as the prediction (i.e., image captioning through retrieval alone). The latter approach yields a CIDEr score of 65.5 on the COCO validation set, substantially lower than the 117.3 from SMALLCAP.



● a child brushing her hair in a room
● a little girl in a dress uses a hair brush on her short hair
● a young girl combs her hair with a yellow comb
● a small girl brushes her hair in front of a mirror'

► **a little girl brushing her hair in front of a mirror**

ø **a little girl holding a cell phone to her ear**

● a small monkey sitting on a stone sill eating a banana
● a young child with an orange piece around their head
● a small brown monkey sitting down while holding a banana
● a small brown monkey holding a ripe banana

► **a small monkey sitting on a rock eating a banana**

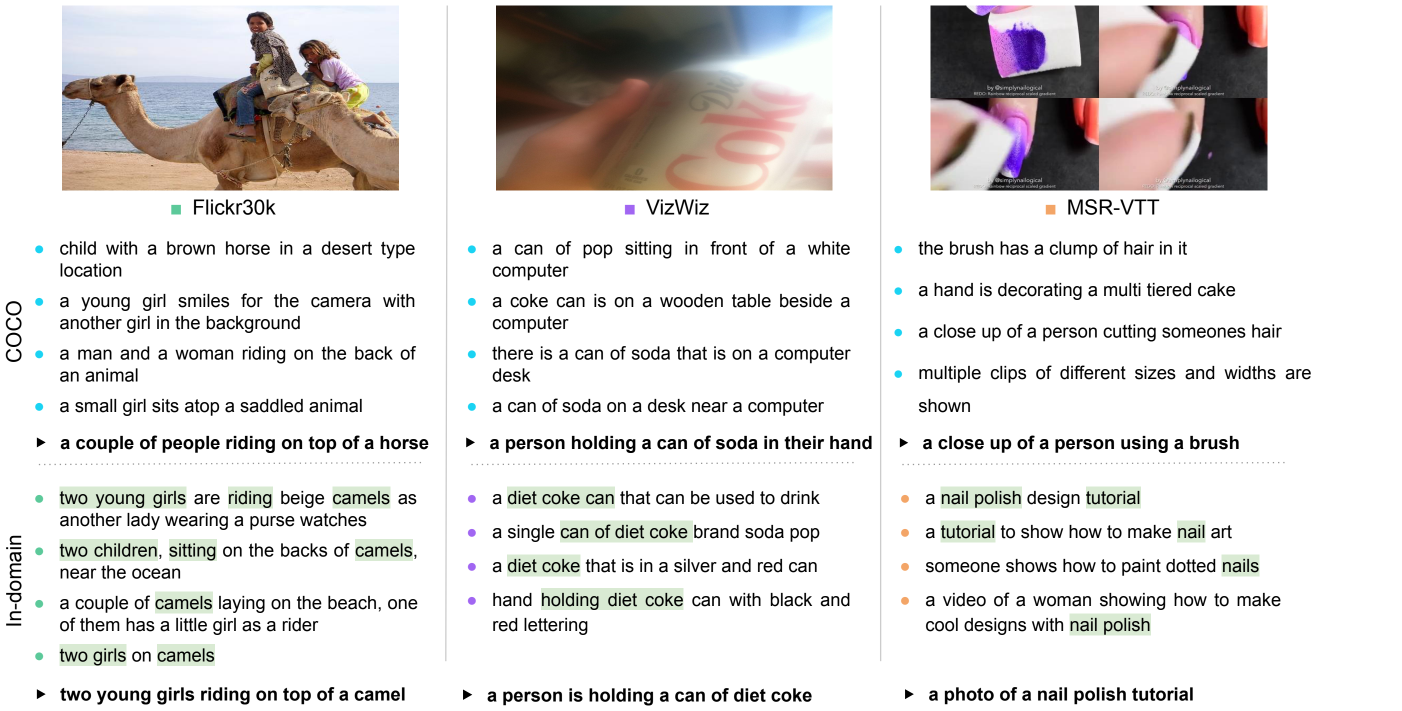
ø **a brown bear eating a piece of food**

● a two dogs in the aisle of a bus next to a bicycle
● a person sitting on a seat on a bus next to brown dogs
● a labrador retriever sits on a public bus
● a person with a dog is standing on a transit vehicle

► **a group of dogs standing next to a bus**

ø **a group of horses standing next to each other**

Figure 8. Caption examples from COCO generated with and without retrieval augmentation. ● denotes the retrieved captions, ► denotes the generated caption from SMALLCAP; ø denotes the caption generated by a model trained without retrieval augmentation.



🟩 Flickr30k

COCO

- child with a brown horse in a desert type location
- a young girl smiles for the camera with another girl in the background
- a man and a woman riding on the back of an animal
- a small girl sits atop a saddled animal

▶ **a couple of people riding on top of a horse**

In-domain

- two young girls are riding beige camels as another lady wearing a purse watches
- two children, sitting on the backs of camels, near the ocean
- a couple of camels laying on the beach, one of them has a little girl as a rider
- two girls on camels

▶ **two young girls riding on top of a camel**

🟪 VizWiz

- a can of pop sitting in front of a white computer
- a coke can is on a wooden table beside a computer
- there is a can of soda that is on a computer desk
- a can of soda on a desk near a computer

▶ **a person holding a can of soda in their hand**

- a diet coke can that can be used to drink
- a single can of diet coke brand soda pop
- a diet coke that is in a silver and red can
- hand holding diet coke can with black and red lettering

▶ **a person is holding a can of diet coke**

🟧 MSR-VTT

- the brush has a clump of hair in it
- a hand is decorating a multi tiered cake
- a close up of a person cutting someones hair
- multiple clips of different sizes and widths are shown

▶ **a close up of a person using a brush**

- a nail polish design tutorial
- a tutorial to show how to make nail art
- someone shows how to paint dotted nails
- a video of a woman showing how to make cool designs with nail polish

▶ **a photo of a nail polish tutorial**

Figure 9. Captions generated for images from the Flickr30k, VizWiz and MSR-VTT datasets, with retrieval either from COCO or in-domain data. With retrieval from in-domain data, SmallCap is less biased towards very frequent concepts, such as *horse*, *soda*, or *brush*, compared to the correct concepts, respectively *camel*, *diet coke* and *nail polish*.