SMALLCAP: Lightweight Image Captioning Prompted with Retrieval Augmentation

Rita Ramos† Bruno Martins† Desmond Elliott⋆,‡ Yova Kementchedjhieva⋆
†INESC-ID, Instituto Superior Técnico, University of Lisbon
⋆Department of Computer Science, University of Copenhagen
‡Pioneer Center for AI
ritaparadaramos@tecnico.ulisboa.pt

Abstract

Recent advances in image captioning have focused on scaling the data and model size, substantially increasing the cost of pre-training and finetuning. As an alternative to large models, we present SMALLCAP*, which generates a caption conditioned on an input image and related captions retrieved from a datastore. Our model is lightweight and fast to train, as the only learned parameters are in newly introduced cross-attention layers between a pre-trained CLIP encoder and GPT-2 decoder.* SMALLCAP *can transfer to new domains without additional finetuning and can exploit large-scale data in a training-free fashion since the contents of the datastore can be readily replaced. Our experiments show that* SMALLCAP*, trained only on COCO, has competitive performance on this benchmark, and also transfers to other domains without retraining, solely through retrieval from target-domain data. Further improvement is achieved through the training-free exploitation of diverse human-labeled and web data, which proves to be effective for a range of domains, including the* nocaps *benchmark, designed to test generalization to unseen visual concepts.*[1]

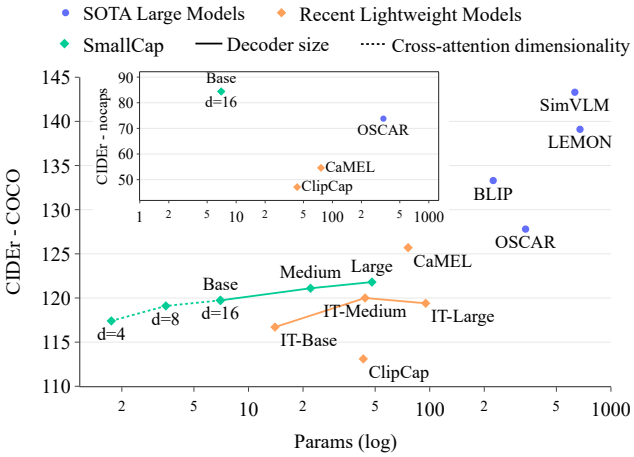

Figure 1. SMALLCAP's performance on the COCO dataset and on the out-of-domain split of the nocaps dataset, compared to other approaches in terms of number of trainable parameters. We can control the number of trainable parameters through the dimensionality of the cross-attention ($d = d_v = d_k$) and the size of the decoder. SMALLCAP is competitive to other lightweight models on COCO, and outperforms much larger models on nocaps.

1. Introduction

The state-of-the-art in image captioning is defined by increasingly large-scale models trained on increasingly large-scale datasets [11, 18, 39, 42]. Scaling up leads to higher computational demands for model pre-training and finetuning on downstream tasks. This becomes especially relevant when numerous model versions may be needed for different visual domains [1] and end-users in practical applications, e.g. image captioning for the visually impaired [10].

Some efforts have been made recently to reduce the cost of model training, e.g., ClipCap [25] and I-Tuning [22]. These models use an off-the-shelf pre-trained vision encoder and language decoder. The parameters of these pre-trained components are frozen and only a mapping between the two is trained for the task of image captioning. This results in a highly reduced number of trainable parameters (∼43M in each case) and faster training time. While these models operate on a much more manageable scale from a research perspective, they can still be unsuitable for the aforementioned practical applications, as both models require separate training for every use-case.

This work presents SMALLCAP, an image captioning model, prompted with captions retrieved from an external datastore of text, based on the input image. This formulation of image captioning enables a range of desirable prop-

[1]Code: https://github.com/RitaRamo/smallcap.

erties: lightweight training, training-free domain transfer, and exploitation of large data in a training-free fashion.

SMALLCAP is both light to train and highly effective (see Figure 1).[2] It uses a pre-trained CLIP vision encoder [29] and GPT-2 language model [31], which are frozen and linked through new cross-attention layers amounting to 7 million trainable parameters. Through retrieval, the model leverages external data and therefore has to store less information within its weights (as demonstrated in Figure 6). Trained on the common COCO benchmark [7], SMALLCAP performs on par with other lightweight-training models, despite an 83% reduction in number of trainable parameters.

SMALLCAP can also leverage data in a training-free manner. Once the model is trained, we can replace the datastore with either (i) captions from a new domain or (ii) a large and diverse collection of captions. In the first case, which presents an alternative to finetuning, SMALLCAP gains access to the style and concepts that characterize the new domain and can generate captions accordingly. In the second case, which presents an alternative to generalized pre-training, SMALLCAP gains access to general knowledge that it can apply to any domain. Our experiments show that SMALLCAP effectively leverages new knowledge accessed through a retrieval-based prompt, improving its performance on different datasets. This includes the challenging VizWiz dataset, where images are captioned for the visually impaired [10], and the `nocaps` challenge dataset with rarely-seen and unseen visual concepts [1].

SMALLCAP competes with other lightweight-training models on in-domain evaluations and outperforms them by a large margin out-of-domain. It overcomes a key limitation of previous models, which require explicit finetuning to adapt to new domains, and in this way attests to the potential of retrieval augmentation for multimodal tasks.

2. Related Work

2.1. Image Captioning Models

Current approaches to image captioning employ encoder-decoder methods, where an input image is passed to a visual encoder and a caption is generated by an autoregressive language decoder [4, 46]. The state-of-the-art is currently held by general purpose vision-and-language (V&L) models [11, 18, 19, 42]. These large-scale models are pre-trained on large amounts of image-text pairs to learn generic multimodal features, after which they can be finetuned to a downstream task such as image captioning, with a separately-optimized model needed for each image captioning dataset. As such, these models require excessive resources for training and deployment.

[2]The `nocaps` results shown in the figure include only models that follow the challenge guidelines, by training on the COCO dataset only.

2.2. Freezing Image Captioning Models

Components of the image captioning model can be initialized with pre-trained weights, frozen in part or completely [2], as a way to prevent catastrophic forgetting [24], i.e. to maintain good generalization. As frozen model parameters require no gradient updates, training becomes faster and occupies less GPU memory. ClipCap and I-Tuning [22, 25] are two lightweight-training image captioning models which use a pre-trained vision encoder, CLIP [29], and language decoder, GPT-2 [31], as frozen model components. To map between these two independently trained components, ClipCap employs prefix-tuning, mapping a fixed-length CLIP embedding of the image into the GPT-2 language space. I-Tuning extracts *visual memory embeddings* from CLIP and uses those to adjust the output hidden states of GPT-2. In SMALLCAP, we also use CLIP and GPT-2, instead connected through a set of trainable cross-attention layers. The novelty here is that SMALLCAP uses retrieval augmentation to maintain performance while substantially reducing the number of trainable parameters.

2.3. Retrieval-Augmented Generation

Retrieval-augmented language generation consists of conditioning generation on additional information that is retrieved from an external datastore [16]. Retrieval augmentation has been gaining traction in other tasks [12, 17], but remains largely unexplored in image captioning. Some relevant works in image captioning include [32–34, 44, 49]. Closest to our work, Sarto *et al*. [34] and Ramos *et al*. [32] recently proposed retrieval-augmented transformer-based captioning models that perform cross-attention over the encoded retrieved captions. Our work differs from previous work in two main ways. We employ a simple prompt-based conditioning method, wherein retrieved captions are used as a prompt to a generative language model. Moreover, we are the first to leverage retrieval augmentation for training-free domain transfer and generalization in image captioning.

2.4. Prompting Text Generation

Prompts have become a common way to pass additional instructions and task demonstrations to a pre-trained language model [30]. In vision-and-language learning, prompts have been used to instruct a model to perform one of multiple tasks it was trained for [18], or to apply it to a new task in a zero-shot fashion [13, 36]. We use prompts with a task demonstration tailored to the specific input image, as a means towards retrieval augmentation.

3. Proposed Approach

3.1. Model

SMALLCAP is a lightweight-training image captioning model augmented with retrieved captions through the use of

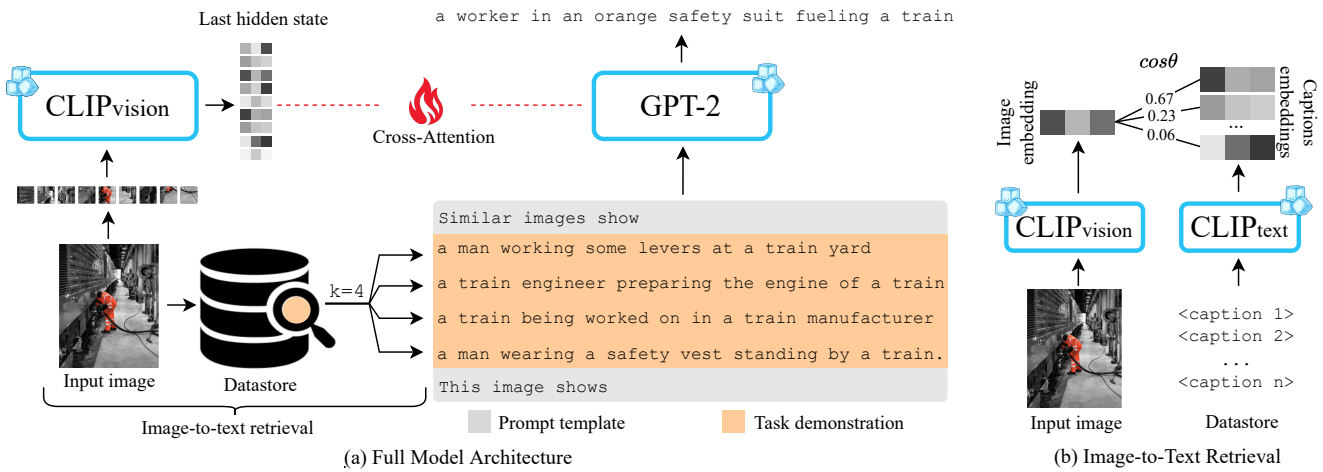

Figure 2. The SMALLCAP approach to image captioning. (a) SMALLCAP generates a caption conditioned on the encoded input image, as well as a set of k retrieved captions which are used as a task demonstration, input to the decoder as a prompt. (b) The k captions are retrieved from a datastore of N captions via image-to-text retrieval.

a prompt. SMALLCAP combines powerful pre-trained unimodal models in an encoder-decoder architecture, as shown in Figure 2 (a). As encoder we use CLIP [29], which produces a sequence of patch embeddings. As decoder we use GPT-2 [31]. These two models operate in different vector spaces, so we connect them with multi-head cross-attention, through which each layer of the decoder attends to the encoder outputs [37]. In order to reduce the compute requirements for training and to preserve their generalization capabilities, we freeze the encoder and decoder and only train the randomly-initialized cross-attention layers between them. We further control the number of trainable parameters through the dimensionality of the projection matrices in the cross-attention layers, which we denote as d. For GPT-2, a model with $d_{model} = 768$ hidden dimensions and $h = 12$ cross-attention heads, d defaults to 64 (d_{model}/h), as per Vaswani *et al.* [37], but can be arbitrarily set to any value (see Appendix A for more details).

Similarly to retrieval-augmented models for other tasks [12, 16, 17, 40], SMALLCAP does not need to store all necessary information within its parameters, because it has access to external knowledge from a datastore of text.

3.2. Prompting with Retrieved Captions

Instead of the image-to-image retrieval methods used in recent work [34], which are limited to image captioning data in the datastore, we employ image-to-text retrieval, as shown in Figure 2 (b). In this way, SMALLCAP can make use of a datastore containing any type of text that is considered useful for describing images, be that image captions, video captions, audio captions, etc. Here, we exploit the full CLIP model, with its vision and text encoders, which map the two modalities into a shared vector space. We encode an input image and the contents of the datastore, and use nearest neighbor search based on cosine similarity to retrieve the k text items from the datastore most similar to the image. The retrieved text is used to fill the slots in a fixed prompt template of the following form: `Similar images show` $\{\texttt{caption}_1\}\ldots\{\texttt{caption}_k\}$. `This image shows ___`.[3] The last sentence of the prompt is similar to the simple, fixed prompts used in other studies [18], but here this cue is preceded by a demonstration of the captioning task, tailored to the input image. The decoder receives this prompt as input tokens and then generates a caption conditioned on the image features $\mathbf{V}$ and the task demonstration $\mathbf{X}$. The weights in the cross-attention layers (θ) are trained by minimizing the cross-entropy loss of predicting the M tokens in the reference $y_1, \ldots, y_M$:

$$L_\theta = -\sum_{i=1}^{M} \log P_\theta(y_i | y_{<i}, \mathbf{X}, \mathbf{V}; \theta). \quad (1)$$

The datastore used to train SMALLCAP can change from training to inference, depending on the application. For example, additional data can be added to enable better generalization, or the datastore can be entirely swapped for new data at inference time to enable domain transfer without the need for retraining, as shown in Section 5.

4. Main Experiments

4.1. Experimental Setup

SMALLCAP's encoder and decoder are initialized respectively from CLIP-ViT-B/32 and GPT-2_{Base}, as available

[3] See Appendix C for more information on the prompt template.

Model	$\|\theta\|$	B@4	M	CIDEr	S
Large Models with V&L pre-training					
LEMON$_{\text{Huge}}$ [11]	675	**41.5**	30.8	139.1	24.1
SimVLM$_{\text{Huge}}$ [42]	632	40.6	**33.7**	**143.3**	**25.4**
OSCAR$_{\text{Large}}$ [19]	338	37.4	30.7	127.8	23.5
BLIP$_{\text{CapFilt-L}}$ [18]	224	39.7	-	133.3	-
Lightweight-training models					
I-Tuning$_{\text{Large}}$ [22]	95	34.8	29.3	119.4	**22.4**
CaMEL [5]	76	**39.1**	**29.4**	**125.7**	22.2
I-Tuning$_{\text{Medium}}$ [22]	44	35.5	28.8	120.0	22.0
ClipCap [25]	43	33.5	27.5	113.1	21.1
I-Tuning$_{\text{Base}}$ [22]	14	34.8	28.3	116.7	21.8
SMALLCAP	**7**	37.0	27.9	119.7	21.3
SMALLCAP$_{\text{d=16, Large}}$	47	37.2	28.3	121.8	21.5
SMALLCAP$_{\text{d=16, Med}}$	22	36.5	28.1	120.7	21.6
SMALLCAP$_{\text{d=8, Base}}$	3.6	36.7	27.8	119.1	21.1
SMALLCAP$_{\text{d=4, Base}}$	1.8	36.0	27.4	117.4	21.0

Table 1. Results on the COCO test set with cross-entropy training. $|\theta|$: number of trainable parameters in the model (in millions).

on HuggingFace [43]. The encoder and decoder are not updated and only the cross-attention layers between them are trained. A 12-head cross-attention layer is added to each of the 12 layers of GPT-2. To achieve a low number of trainable parameters, we vary the dimensionality of the projection matrices in the cross-attention layers, d, by scaling from the default size of 64 down to 16, 8 and 4, which results in model variants with 7M, 3.6M and 1.8M trainable parameters, respectively. Our main model, SMALLCAP, has 7M trainable parameters and a total of 218M parameters (including the frozen CLIP encoder and GPT-2 decoder).

The cross-attention layers are trained on the COCO dataset [7] using the standard Karpathy splits [15]. The models are trained to minimize the cross-entropy loss using an AdamW optimizer [21] with an initial learning rate of 1e-4 and a batch size of 64. Training runs for 10 epochs and we use the epoch checkpoint with the best CIDEr score on the validation set. Training takes up to 8 hours on a single NVIDIA A100 GPU, using 16 GB of the available memory.

During training, the model is prompted with a set of $k = 4$ captions per image, retrieved from a datastore of the training captions from COCO. Retrieval is based on CLIP-ResNet-50x64[4] representations of input images and captions in the datastore, the latter being precomputed offline and indexed with FAISS [14] for efficient nearest neighbor searching.[5] During inference, the model generates a caption using beam search decoding with a beam size of 3. Inference, including retrieval and prompting, takes 0.22 seconds on average across 1,000 randomly sampled images, compared to 0.19 seconds without retrieval. For more details on design choices and hyperparameters, see Appendix B.

For evaluation, we compute the standard metrics: BLEU-4 (B@4) [27], METEOR (M) [8], CIDEr [38], and SPICE (S) [3], using the COCO evaluation package.[6]

[4]Downloaded from https://github.com/openai/CLIP

[5]We use an inner product index (`IndexFlatIP`) without any training and normalize the representations to search based on cosine similarity.

Model	In	Near	Out	Entire
OSCAR$_{\text{Large}}^{\diamond}$	84.8	82.1	73.8	80.9
CaMEL*	**88.1**	79.1	54.6	75.9
ClipCap*	74.5	65.6	47.1	63.4
SMALLCAP	83.3	77.1	65.0	75.8
SMALLCAP$_{\text{+W+H}}$	87.9	**84.6**	**84.4**	**85.0**

Table 2. CIDEr results on the `nocaps` test set. $\diamond$: Results copied from the respective publications. $\star$: Results computed by us. +W+H: datastore with additional Web and Human-labeled data.

4.2. Benchmark Results

Here, we report results on COCO [7], as well as on `nocaps` [1], a challenge dataset for evaluating the generalization capabilities of models trained on COCO.

COCO: In Table 1 we benchmark our approach on the COCO dataset. In the top half of the table, we acknowledge the strong performance of large-scale pre-trained models, ranging in size from 224M to 675M trainable parameters. We also note that these models are pre-trained on 4M–1.8B image-caption pairs, i.e., much more than the COCO data.

In the lower half of the table we see how our approach compares to other lightweight-training models. With only 7M parameters, SMALLCAP performs better or on par with ClipCap and I-Tuning. In this in-domain setting, it is only outperformed by CaMEL, which is trained end-to-end with eleven times as many trainable parameters. Reducing the number of trainable parameters to 3.6M, SMALLCAP$_{\text{d=8, Base}}$ still yields competitive performance, and even with just 1.8M trainable parameters, SMALLCAP$_{\text{d=4, Base}}$ is better than the substantially larger models ClipCap and I-Tuning$_{\text{Base}}$. We also experiment with Medium and Large GPT-2 decoders (SMALLCAP$_{\text{Medium}}$ and SMALLCAP$_{\text{Large}}$ in Table 1), and find that performance scales: by one CIDEr point from Base to Medium and by another point from Medium to Large.[7] Despite its small size, SMALLCAP shows competitive performance on COCO, the dataset it was trained on. In contrast to previous lightweight-training models, SMALLCAP further has the ability to generalize and transfer out-of-domain without retraining, as shown in subsequent experiments.

[6]https://github.com/tylin/coco-caption

[7]See Appendix E for more results regarding scaling the decoder.

nocaps: Results on the nocaps test set are reported in Table 2.[8,9,10] SMALLCAP clearly outperforms other lightweight methods *Out*-of-domain and achieves competitive performance *In*-domain and *Near*-domain. The model's strong generalization capabilities point to it being less prone to over-fitting as it does not need to memorize its training data, available also through retrieval. Our model can further improve when additional data is placed in the datastore, as seen in SMALLCAP$_{+W+H}$. In this variant, described in more detail in Section 5.2, the COCO datastore is augmented with diverse web (W) and human-labeled (H) data. SMALLCAP$_{+W+H}$ shows impressive generalization capabilities, outperforming the much larger OSCAR$_{Large}$ by over 10 points in the *Out*-of-domain setting. Next, we further explore SMALLCAP's ability for training-free transfer to new domains on diverse datasets.

5. Training-Free Use of Data

In this section, we study SMALLCAP's ability to leverage new data in its datastore in a training-free manner, i.e. all experiments presented here constitute changes made to the datastore at inference time, while the model, trained on COCO, remains fixed. The focus is on out-of-domain performance as measured on a diverse set of captioning datasets: Flick30k [47], VizWiz [10] and MSR-VTT [45]. The latter is in fact a video captioning dataset, which we adapt by converting video clips into an image of four 4 frames, sampled at 0, 25, 50 and 100% of the clip duration (see the MSR-VTT example in Figure 5). We start by exploring different configurations of the datastore, with the results in Table 3 reported on validation data.

5.1. In-domain Data

In the top of Table 3, we show how SMALLCAP performs when its datastore is populated with the training data associated with each respective dataset (*In-domain*). In comparison to using COCO captions in the datastore (*COCO*), the model performance substantially increases for all three datasets. This shows that SMALLCAP adapts to the retrieved information to achieve domain transfer. The improvement is most notable for VizWiz, likely because the nature of this dataset is very distinct from COCO, and thus there is a larger domain gap to be closed.

5.2. Augmenting the Datastore

In Table 3 (*Datastore augmentation*), we augment the in-domain datastore with additional large-scale data in an effort to improve generalization. We experiment with diverse web data (which is large-scale but automatically labeled) and human-labeled data (smaller-scale but clean).[11]

SMALLCAP datastore	F30K	VW	MV
COCO	52.2	34.5	23.3
In-domain	55.4	47.7	29.2
Datastore augmentation			
In-domain + **W**eb	**58.6**	**48.0**	29.8
In-domain + **H**uman-labeled	57.6	47.5	**30.9**
In-domain + **W** + **H**	57.9	**48.0**	30.7
Domain-agnostic			
Web	58.4	42.4	27.6
Human-labeled	56.6	36.4	29.0
Web + Human-labeled	57.8	42.2	29.9

Table 3. Exploration of the training-free use of data. Validation performance of SMALLCAP measured in CIDEr score, with different contents of the datastore, without any finetuning on Flickr30k (F30K), VizWiz (VW), and MSR-VTT (MV). The best number per section is underlined; the best number overall is in bold.

+ Web Data: We first consider large-scale data from the web, expanding the datastore with text from three web datasets [18] (Conceptual Captions [35], Conceptual 12M [6], and SBU captions [26]).[12] The results with *In-domain + Web* in Table 3 show that performance improves for all three datasets. We can see a bigger improvement on Flickr30K and MSR-VTT when using a large and diverse datastore compared to just using in-domain data. Improvement on VizWiz, on the other hand, remains low, in line with the earlier observation that this dataset has a distinct distribution that is not easily matched by other data.

+ Human-labeled Data: We also consider smaller-scale but clean human-labeled data. As discussed in Section 3.2, the datastore can contain any type of text that can be useful to describe images, thus not being constrained by the assumption of image-caption pairs. As such, we consider text not only from image captions (COCO [7], Flickr30k [47], VizWiz [10]), but also from video captions (MSR-VTT [45], VATEX [41], TGIF [20]), audio captions (Clotho [9]), and localized narratives (LN ADE20k, LN COCO, LN Flickr30k, LN OpenImages [28]).

As seen in *In-domain + Human-labeled*, adding human-labeled data to the datastore leads to an improvement over using in-domain data only for Flickr30k and MSR-VTT but not for VizWiz. In comparison to *In-domain + Web*, this

[8] OSCAR$_{Large}$ results with COCO-only training. CaMEL results with CLIP-ResNet-50×16, $\lambda_k d$ = 0.1, no mesh connectivity, and a cross-entropy objective (checkpoint obtained through personal communication).

[9] We only include results from models which follow the nocaps guidelines to not train on image-caption pairs beyond COCO [1]. As we use only captions for retrieval, our method is also in line with these guidelines.

[10] We also include results on the validation set in Appendix D.

[11] Data size and further details can be found in Appendix F.

[12] We use a trained FAISS index (IndexIVFFlat) for faster search.

	Flickr30K	VizWiz	MSR-VTT
ClipCap	41.2	28.3	12.5
CaMEL	55.2	37.6	20.7
SMALLCAP	**60.6**	**55.0**	**28.4**
Pre-training & finetuning			
SOTA	79.6 [23]	120.8 [39]	75.9 [39]

Table 4. Out-of-domain performance without additional training, measured in CIDEr score on the test data. Flickr30K and VizWiz results with *In-domain* + *Web*, and MSR-VTT result with *In-domain* + *Human-labeled*. We include SOTA results from large-scale pre-trained models, finetuned on the respective datasets.

improvement is smaller for Flickr30k, but larger for MSR-VTT. Although smaller than web data, human-labeled data benefits MSR-VTT more, because it contains text from different tasks, including video captioning.

+ Web + Human-labeled Data: Seeing that SMALLCAP can benefit both from Web and from Human-labeled data as augmentations over in-domain data alone, we also consider a combination of the two, to determine whether their contributions are complementary or overlapping. The results for *In-domain* + *W* + *H* in Table 3 show that combining the two sources of data is not beneficial for any of the three datasets.

5.3. Domain-agnostic Datastore

In this section, we study whether SMALLCAP could still perform well without access to in-domain data and report results under the heading *Domain-agnostic* in Table 3. We find that the patterns observed above with in-domain data largely hold without it as well. With the large and diverse *Web* datastore, SMALLCAP performs close to or even better than with *In-domain* data. *Human-labeled* data is again seen to benefit MSR-VTT the most, the optimal configuration for this dataset being *Web* + *Human-labeled*.

From the exploration presented above, we conclude that SMALLCAP's image captioning capabilities can transfer with access to web data in addition to or in place of in-domain data. The model can also leverage human-labeled data beyond image-captioning pairs in solving tasks other than image captioning, such as video captioning.

5.4. Results with the Best Configuration

Having explored different datastore configurations for each of the three datasets, we use the best configuration for each to compare zero-shot performance against ClipCap and CaMEL, both models also trained only on COCO. In Table 4 we show test set performance (in CIDEr score) with a datastore consisting of *In-domain* + *Web* for Flickr30k and VizWiz, and *In-domain* + *Human-labeled* for MSR-VTT. SMALLCAP outperforms both ClipCap and CaMEL by a large margin on all three datasets. In comparison to CaMEL, the stronger baseline of the two, we see a 5.4 point improvement on Flickr30k, a noteworthy 17.4 point improvement on VizWiz and an increase of 7.7 points on MSR-VTT. The large improvement on VizWiz demonstrates SMALLCAP's ability to transfer to domains very distinct from the training data, i.e., COCO. The improvement on MSR-VTT, on the other hand, shows our approach has potential not only for other domains but for other tasks as well. These results show that while other lightweight-training models lack out-of-domain generalization without finetuning, our model can transfer across domains by only swapping the datastore contents. In the bottom of the table, we provide state-of-the-art results for context, which were achieved by large-scale pre-trained V&L models, finetuned specifically on the respective datasets.

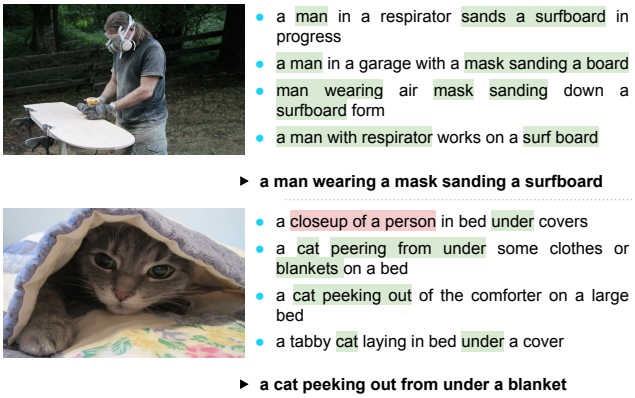

Figure 3. Examples generated by SMALLCAP, together with the retrieved predictions from the COCO datastore. • denotes the retrieved captions, highlighted as green or red to indicate correct and mismatch captions, respectively. ▸ denotes the generated caption.

6. Discussion

6.1. Qualitative Examples

Figure 3 shows examples of the retrieved and generated captions for two images from the COCO dataset. In first example, we observe that the retrieved captions are highly relevant to the input image and the generated captions are semantically similar to them. As seen in the second example, SMALLCAP can also be robust to misleading information from retrieval. Figure 5 shows examples of captions generated for Flickr30k, VizWiz, and MSR-VTT, with a datastore populated with COCO or with in-domain data. These qualitative results show how SMALLCAP adapts to new domains: with the help of the retrieved captions, it correctly refers to the concepts *tutu*, the *Swanson* brand name, and *Pokemon*. The first two concepts are not present in the COCO training data at all, while the last is seen just six times.

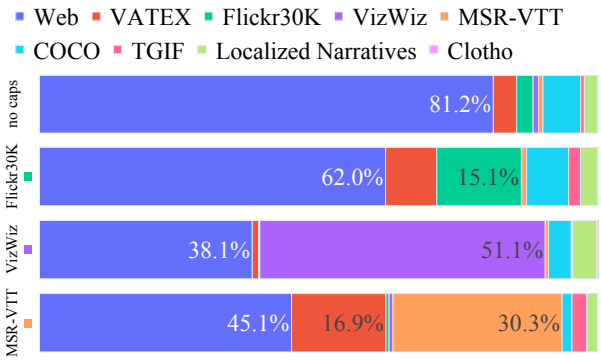

Figure 4. Percentage of the retrieved captions that come from each data source, when testing the model on the different benchmarks of `nocaps` , Flickr30k, VizWiz and MSR-VTT.

6.2. Analysis of the Retrieved Captions

In Section 5.3, we demonstrated the ability of SMALLCAP to exploit large data in a training-free fashion. Here, we inspect the distribution of retrieved captions in the *In-domain + Web + Human-labeled* setting, in order to understand the individual impact of each dataset. As can be seen in Figure 4, most text is retrieved from web data, especially in the presence of unseen visual concepts, as is the case for `nocaps`. Besides web data, the model tends to retrieve text from the corresponding dataset or from a similar domain; for instance, MSR-VTT retrieval also relies on other video datasets. Due to its unique distribution, VizWiz stands out as the case with the highest rate of in-domain retrieval.

Seeing that text from all types of human-labeled data is retrieved, we measure the actual impact of each type on performance. In Table 5, we report performance on Flickr30k, VizWiz, and MSR-VTT, with an in-domain datastore augmented with either Image captions, Video captions, localized Narratives, or Audio captions. We see that SMALLCAP can indeed benefit from data beyond image captions. For instance, video captions help not only for MSR-VTT, but also for Flickr30k and VizWiz. Flickr30k benefits the most from localized narratives since this dataset contains narratives for the Flickr30k images. Audio captions are beneficial for both Flickr30k and MSR-VTT. Considering the distinct nature of the audio and visual modalities, this finding demonstrates the potential of leveraging data which has previously seen limited application to image captioning.

6.3. The Impact of Retrieval

In Figure 6, we show validation performance with 1.8, 3.6, 7, 14 and 28 million trainable parameters with and without retrieval augmentation.[13] For variants with retrieval augmentation, performance is stable across the range of model sizes considered. Reducing the number of trainable parameters by a factor of four, from 28M to 7M, leads to a slight drop of 0.6 CIDEr points. This indicates that SMALLCAP has a close-to-optimal size to performance trade-off.

Next, we ablate the retrieval augmentation to quantify its impact. We train models without retrieval augmentation, prompting them with just the phrase `This image shows`. As seen in Figure 6, without the aid of retrieved captions, there is a notable drop in performance compared to results with retrieval. Moreover, model performance degrades at a higher rate: while performance at the two extremes of model sizes differs by just 1.7 CIDEr points with retrieval, without it the difference is 4.3 points.[14]

In order to confirm that SMALLCAP is not simply paraphrasing the retrieved captions without attending to the visual input, we experiment with ablating the visual modality. For this, we train a model on "blank" input images, setting the visual features from the encoder to zero. This yields a much lower CIDEr score of 90.1 on the validation set, showing that SMALLCAP indeed uses the visual input.

[13] The model sizes correspond to $d = 4, 8, 16, 32$ and 64.

	Flickr30K	VizWiz	MSR-VTT
In-domain	52.2	47.7	29.2
+ Image	56.7	47.8	29.8
+ Video	57.0	47.8	31.1
+ Narratives	57.1	47.2	28.7
+ Audio	55.4	47.7	29.4

Table 5. SMALLCAP performance with retrieval from the different sources of the Human-labeled data. The model can benefit from having access to text that is not only from image captioning tasks, but also from other tasks such as audio captioning.

Decoder	$\|\theta\|$	B@4	M	CIDEr	S
GPT2-Base$_{d=16}$	7	37.0	27.9	119.7	21.3
OPT-125M$_{d=16}$	7	37.6	28.4	122.0	21.7
GPT2-Medium$_{d=16}$	22	36.5	28.1	120.7	21.6
OPT-350M$_{d=16}$	22	37.5	28.7	122.7	22.0

Table 6. Results with different decoders on the COCO test set.

6.4. Alternative Decoders

At the request of the anonymous reviewers, we include additional experiments with some more recent language models: OPT-125M and OPT-350M [48], equivalent in size to GPT2-Base and GPT2-Medium.[15] The results in Table 6 show that our approach performs well with these stronger

[14] See Appendix H for qualitative examples with and without retrieval.

[15] There is no OPT variant equivalent in size to GPT2-Large.

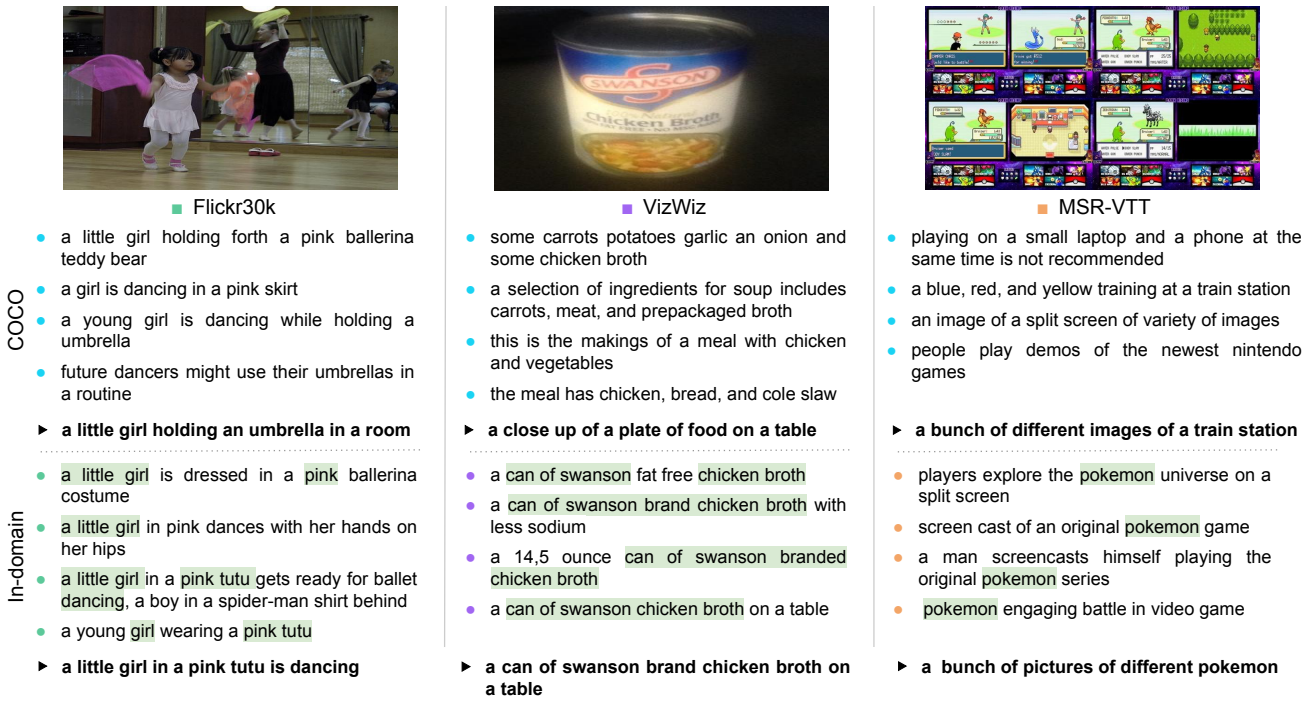

Figure 5. Examples of captions generated for Flickr30k, VizWiz and MSR-VTT, with retrieval either from COCO or in-domain data. The captions use words retrieved from the in-domain datastores which were rarely seen in the COCO training data (tutu, swanson, pokemon).

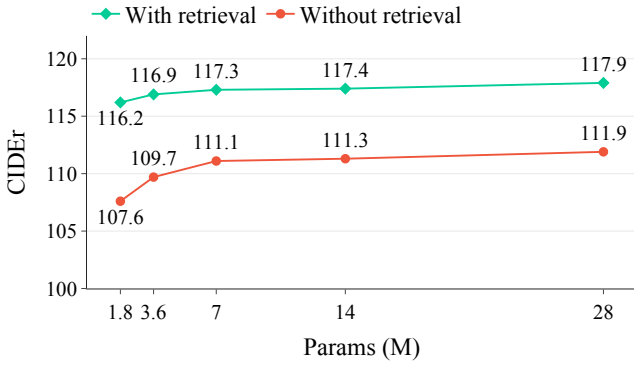

Figure 6. CIDEr scores on the COCO validation set, with and without retrieval, across different cross-attention sizes.

language models and is therefore model agnostic.[16,17]

7. Conclusion

In this paper, we propose SMALLCAP, an image captioning model augmented with retrieval, which is light to train and can be transferred across domains without retraining. Results on the COCO dataset show that SMALLCAP is competitive to other lightweight-training models despite having substantially less trainable parameters, instead leveraging non-parametric information from a datastore of text. Out-of-domain evaluations show that SMALLCAP can also perform training-free domain transfer when given access to a datastore with target-domain data. Our model further benefits from diverse web and human-labeled data in addition to or in place of target-domain data. We find that SMALLCAP benefits not just from access to image captions, but also to video and audio captions (resources neglected in image captioning work in the past).

SMALLCAP's small size and impressive performance in out-of-domain settings attest to the potential of retrieval augmentation as an alternative to the expensive training found in large pre-trained vision-and-language models and the costly finetuning that even previous lightweight-training models require in order to adapt to different image captioning datasets. Future work can apply our retrieval augmentation approach to a wider range of multimodal tasks, and further explore the scalability of the data used for retrieval.

[16] See Appendix E for OPT results without retrieval.

[17] Due to our academic computing budget, we only repeat the experiments from Table 1. Future work can experiment further in this direction.

Acknowledgements

This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, through Fundação para a Ciência e Tecnologia (FCT) with the Ph.D. scholarship 2020.06106.BD, and through the INESC-ID multi-annual funding from the PIDDAC programme (UIDB/50021/2020).

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1, 2, 4, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 4

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[5] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. CaMEL: Mean Teacher Learning for Image Captioning. In *International Conference on Pattern Recognition*, 2022. 4

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 4, 5, 12

[8] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 4

[9] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020. 5, 12

[10] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer, 2020. 1, 2, 5, 12

[11] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 1, 2, 4

[12] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022. 2, 3

[13] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021. 2, 11

[14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 4

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 4

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. 2, 3

[17] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022. 2, 3

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2, 3, 4, 5, 12

[19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2, 4, 11

[20] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 5, 12

[21] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 4

[22] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. I-tuning: Tuning language models with image for caption generation. *arXiv preprint arXiv:2202.06574*, 2022. 1, 2, 4

[23] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. Vc-gpt: Visual conditioned gpt for end-to-end generative vision-and-language pre-training. *arXiv preprint arXiv:2201.12723*, 2022. 6

[24] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[25] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021. 1, 2, 4

[26] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 5

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[28] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020. 5, 12

[29] 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, 3

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 3

[32] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*, 2023. 2

[33] Rita Parada Ramos, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, and Bruno Martins. Retrieval augmentation for deep neural networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[34] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. *arXiv preprint arXiv:2207.13162*, 2022. 2, 3

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[36] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc., 2021. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3, 11

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 4

[39] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1, 6

[40] Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179, Dublin, Ireland, May 2022. Association for Computational Linguistics. 3

[41] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 5, 12

[42] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 2, 4

[43] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. 4

[44] Chunpu Xu, Wei Zhao, Min Yang, Xiang Ao, Wangrong Cheng, and Jinwen Tian. A unified generation-retrieval framework for image captioning. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019. 2

[45] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5, 12

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[47] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5, 12

[48] 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. 7, 12

[49] Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang, and Jiaxuan Zhang. Image caption generation via unified retrieval and generation-based method. *Applied Sciences*, 10(18), 2020. 2