This is the peer reviewd version of the followng article:

Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets / Cornia, Marcella; Baraldi, Lorenzo; Fiameni, Giuseppe; Cucchiara, Rita. - In: INTERNATIONAL JOURNAL OF COMPUTER VISION. - ISSN 0920-5691. - 132:5(2024), pp. 1701-1720. [10.1007/s11263-023-01949-w]

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

05/03/2026 22:35

(Article begins on next page)

Generating More Pertinent Captions by Leveraging Semantics and Style on Multi-Source Datasets

Marcella Cornia[1†], Lorenzo Baraldi[1†], Giuseppe Fiameni[2], Rita Cucchiara[1,3]

[1*]University of Modena and Reggio Emilia, Modena, Italy.
[2]NVIDIA AI Technology Centre, Bologna, Italy.
[3]IIT-CNR, Pisa, Italy.

*Corresponding author(s). E-mail(s): marcella.cornia@unimore.it;
Contributing authors: lorenzo.baraldi@unimore.it; gfiameni@nvidia.com;
rita.cucchiara@unimore.it;
[†]These authors contributed equally to this work.

Abstract

This paper addresses the task of generating fluent descriptions by training on a non-uniform combination of data sources, containing both human-annotated and web-collected captions. Large-scale datasets with noisy image-text pairs, indeed, provide a sub-optimal source of supervision because of their low-quality descriptive style, while human-annotated datasets are cleaner but smaller in scale. To get the best of both worlds, we propose to leverage and separate semantics and descriptive style through the incorporation of a style token and keywords extracted through a retrieval component. The proposed model avoids the need of object detectors, is trained with a single objective of prompt language modeling, and can replicate the style of human-collected captions while training on sources with different input styles. Experimentally, the model shows a strong capability of recognizing real-world concepts and producing high-quality captions. Extensive experiments are performed on different image captioning datasets, including CC3M, nocaps, and the competitive COCO dataset, where our model consistently outperforms baselines and state-of-the-art approaches.

Keywords: Image Captioning, Vision and Language, Multimodal Learning

1 Introduction

Image captioning, which aims at generating textual descriptions from visual inputs, has emerged as an attractive research problem in the last few years, as it entails modeling the connections between the visual and textual modalities (X. Li et al., 2020; P. Zhang et al., 2021) and can be seen as a fundamental step toward machine intelligence. This has led to the development of effective strategies for feature extraction (Anderson et al., 2018), cross-modal modeling (Cornia, Stefanini, Baraldi, & Cucchiara, 2020; Pan, Yao, Li, & Mei, 2020) and model training (X. Li et al., 2020; Rennie, Marcheret, Mroueh, Ross, & Goel, 2017).

Despite these advances, researchers are still working to endow such models with the ability to describe the entire variety of real-world concepts. The issue boils down to the limitations of popularly-used human-annotated datasets like COCO (Gurari, Zhao, Zhang, & Bhattacharya, 2020; T.-Y. Lin et al., 2014; Young, Lai, Hodosh,

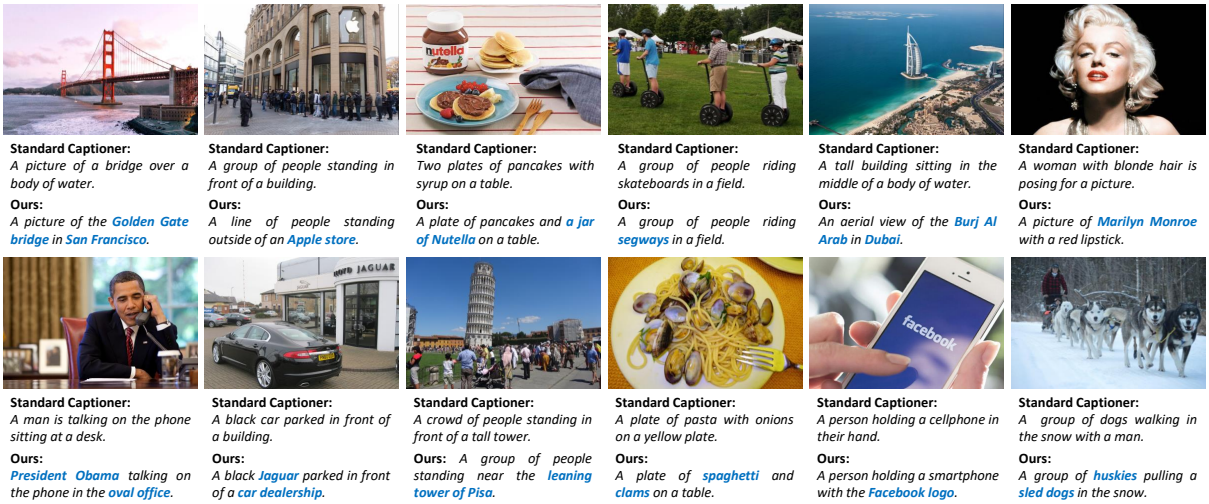


Fig. 1 Sample descriptions generated by our model, in comparison with a Transformer-based captioner trained on COCO. Our approach generates high-quality captions by separating content from style.

& Hockenmaier, 2014) which, while being high-quality in terms of descriptive style, are limited in terms of semantic variability and size. Recent efforts have automatically collected large-scale datasets with noisy image-text pairs from the web, partially solving the semantic scale issue at the cost of reducing the quality of the annotations (Changpinyo, Sharma, Ding, & Soricut, 2021; Ordonez, Kulkarni, & Berg, 2011; Schuhmann et al., 2022; Sharma, Ding, Goodman, & Soricut, 2018; Srinivasan, Raman, Chen, Bendersky, & Najork, 2021).

In this work we focus on generating captions that can be richer in terms of semantics and include proper names and long-tail concepts (Fig. 1), thus being more pertinent to the input image. We do this by jointly leveraging web-collected and human-annotated sources and maintaining the style and fluency of human-annotated captions. The core idea behind our approach is that of separating semantics and descriptive style while training on non-homogeneous data sources. This is achieved through the introduction of a *style token* that can condition the network both at training and generation time. During training, the token is employed to distinguish between human-annotated and web-crawled sources. At generation time, the style token can be used to generate descriptions which resemble the style of human-annotated ones, enriched by the semantics learned on web-collected datasets (Fig. 2).

Further, to better represent semantics, we extract textual keywords through a novel retrieval-based approach (Radford et al., 2021), which avoids the need of using tags or descriptions from object detectors (Anderson et al., 2018; P. Zhang et al., 2021). This also allows us to scale beyond a limited set of categories and fully represent the semantics of the image regardless of its source. The addition of the style token and of textual keywords fosters the transfer of descriptive style and semantic concepts between data sources.

From the point of view of the architecture, our model features an encoder-decoder structure that clearly separates the visual and textual domain, in contrast to the paradigm of employing BERT-like architectures (X. Li et al., 2020; Zhou et al., 2020). To represent images, we employ a multimodal feature extractor based on CLIP (Dosovitskiy et al., 2021; Radford et al., 2021) which can directly take raw pixels as input and avoids the need of using object detectors. Finally, our model is trained using only a language modeling objective and does not require complex pre-training strategies.

Our model outperforms existing proposals in terms of caption quality, sometimes also surpassing models trained on significantly larger datasets (Wang et al., 2022), and shows an improved capability of generating named entities to improve the description pertinence. Experimentally, we assess the performance of the proposed approach on different image captioning

datasets, including COCO (T.-Y. Lin et al., 2014), nocaps (Agrawal et al., 2019), VizWiz (Gurari et al., 2020), and TextCaps (Sidorov, Hu, Rohrbach, & Singh, 2020) that all contain human-annotated sentences, and CC3M (Sharma et al., 2018), WIT (Srinivasan et al., 2021), and a portion of LAION-400M (Schuhmann et al., 2021) that instead are composed of web-collected data. Overall, our work demonstrates that heterogeneous data sources can be properly exploited, together with a selective architecture, to increase the performance of image captioning systems.

Contributions. To sum up, the contributions of this paper are fourfold:

- We propose a framework for learning on non-uniform collections of caption sources while maintaining a separation between semantic content and descriptive style. This allows the generation of fluent descriptions that resemble the quality of human-collected ones while learning from web-scale data.
- Our approach employs a style token as a means to separate the descriptive styles of human-annotated and web-collected sources, and textual keywords extracted through a retrieval component to represent semantics.
- In terms of architecture, our model features a fully-attentive encoder-decoder structure that jointly encodes keywords, style, and text, and is trained with a single objective of prompt language modeling.
- We evaluate our model against carefully-designed baselines and recent approaches on COCO, nocaps, VizWiz, TextCaps, CC3M, WIT, and LAION-400M. On COCO, our approach reaches a performance of 149.6 CIDEr points.

2 Related Work

Image Captioning. Research on image captioning has jointly focused on modeling the visual encoding pipeline, the language model, and the multi-modal connections between them (Stefanini et al., 2022). While traditional approaches have focused on training on curated datasets, the recently emerged pre-training paradigm (Y.-C. Chen et al., 2020; Hu et al., 2020; Lu, Batra, Parikh, & Lee, 2019; Tan & Bansal, 2019; Zhou et al., 2020) aims at learning from weakly

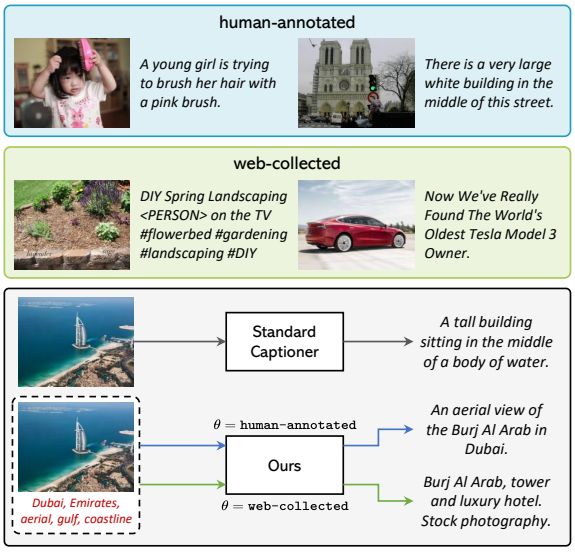


Fig. 2 Samples of human-annotated and web-collected (image, caption) pairs and overview of our approach.

labeled or noisy sources. Most of the approaches have employed BERT-like (Devlin, Chang, Lee, & Toutanova, 2018) or encoder-decoder architectures in conjunction with self-supervised or sequence learning objectives. The OSCAR (X. Li et al., 2020) model considers triplets of object tags, detections, and captions and trains using a combination of masked token loss and contrastive loss. VinVL (P. Zhang et al., 2021) employs the same objectives while proposing a better object detector, and trains on 8.85 million text-image pairs. SimVLM (Wang et al., 2022), instead, uses an encoder-decoder architecture and learns visual features from scratch, training on the large ALIGN dataset (Jia et al., 2021). Recent models like LEMON (Hu et al., 2022) and BLIP (J. Li, Li, Xiong, & Hoi, 2022) have further investigated the scaling properties of captioning models, also adopting custom architectures.

Recently, following the latest trends of large-scale language models (Brown et al., 2020; S. Zhang et al., 2022), several large-scale multi-modal solutions have been proposed (Alayrac et al., 2022; J. Li, Li, Savarese, & Hoi, 2023; Yu et al., 2022), considerably increasing the number of parameters up to few billion (*e.g.* the well-known Flamingo model (Alayrac et al., 2022) has 10.6B parameters) and, consequently, the computational complexity.

Visual encoders. In terms of visual encoding, after the emergence of global (Karpathy & Fei-Fei, 2015; Rennie et al., 2017) and grid descriptors (K. Xu et al., 2015), the use of object detections (Anderson et al., 2018; P. Zhang et al., 2021) has become one of the most popular approaches. Indeed, it provides clean visual elements and a partial bridge between the visual and the textual domains. While several works have encoded regions through graph-based (Yang, Tang, Zhang, & Cai, 2019) or self-attentive structures (Cornia et al., 2020; Huang, Wang, Chen, & Wei, 2019; Pan et al., 2020), the emergence of self-attentive visual encoders (Dosovitskiy et al., 2021) and large-scale multi-modal models has enabled new strategies, ranging from training better detectors to having end-to-end visual models trained from scratch (Kim, Son, & Kim, 2021; H. Xu et al., 2021; Yan et al., 2021; P. Zhang et al., 2021). Recently, Shen et al. (2022) showed that features encoded by large-scale multi-modal architectures like CLIP (Radford et al., 2021) perform at least on par with detection-based approaches. These findings have been confirmed by subsequent methods (Barraco, Cornia, Cascianelli, Baraldi, & Cucchiara, 2022; Mokady, Hertz, & Bermano, 2021), also augmenting the captioning model with either knowledge distillation (Barraco, Stefanini, et al., 2022) or retrieval components (Y. Li, Pan, Yao, & Mei, 2022; Sarto, Cornia, Baraldi, & Cucchiara, 2022).

Semantic content and style separation. A related line of research focuses on the generation of textual descriptions conditioned on a specific style or sentiment. In this context, some solutions have been proposed to train a captioning model with image-text pairs composed of images directly associated with positive and negative sentences (Mathews, Xie, & He, 2016), eventually constraining the captioning model to also preserve the ability to generate factual descriptions (*i.e.* captions without a specific style) (T. Chen et al., 2018). To reduce the dependency on paired data, other approaches proposed to leverage unpaired stylized corpus to generate more accurate stylized captions (Gan, Gan, He, Gao, & Deng, 2017; Guo, Liu, Yao, Li, & Lu, 2019; Mathews, Xie, & He, 2018; Zhao, Wu, & Zhang, 2020). In particular, Guo et al. (2019) employed a one-hot style indicator to condition their captioning model, which resembles our style token. Differently, the work introduced by (Klein, Mahajan, & Roth, 2021) also considers the stylistic content of the input image represented with style-specific attributes.

While these solutions can provide a good strategy to personalize the generation of a captioning model thus potentially meeting the requirements of different users, only a few attempts have been done on the separations of semantics and style. For example, recent works on dataset collection like Conceptual Captions (Changpinyo et al., 2021; Sharma et al., 2018) have proposed strategies for filtering lower quality captions to ensure minimum quality levels. BLIP (J. Li et al., 2022) has explicitly addressed quality assurance in web-scale training by proposing a learnable image-grounded caption filter. To our knowledge, we are the first to address the separation of semantics and style as a method for dealing with noisy web-collected captions.

3 Proposed Method

The goal of an image captioning algorithm is that of modeling an autoregressive probability distribution over words conditioned on an input image, *i.e.*

$$p(\boldsymbol{w}_t | \boldsymbol{w}_{\tau<t}, \mathbf{V}), \tag{1}$$

where $\mathbf{V}$ represents an input image and $\{\boldsymbol{w}_t\}_t$ is the sequence of words comprising the generated caption. In previous works, this is usually achieved by training a language model conditioned on visual features to mimic ground-truth descriptions.

The relation between images and word sequences, though, is far from being bijective, and the same semantic content can be described in different ways according to intent and descriptive style. The variance in descriptive style is a key element that differentiates human-annotated captions from noisy web-collected datasets. This is well testified in Fig. 2, which compares web-collected caption sources like Conceptual Captions (Sharma et al., 2018) or YFCC100M (Thomee et al., 2016) and human-annotated sources like COCO. The latter feature a grammatically correct, constant, and generic descriptive style. The former, instead, can have heterogeneous descriptive styles depending on

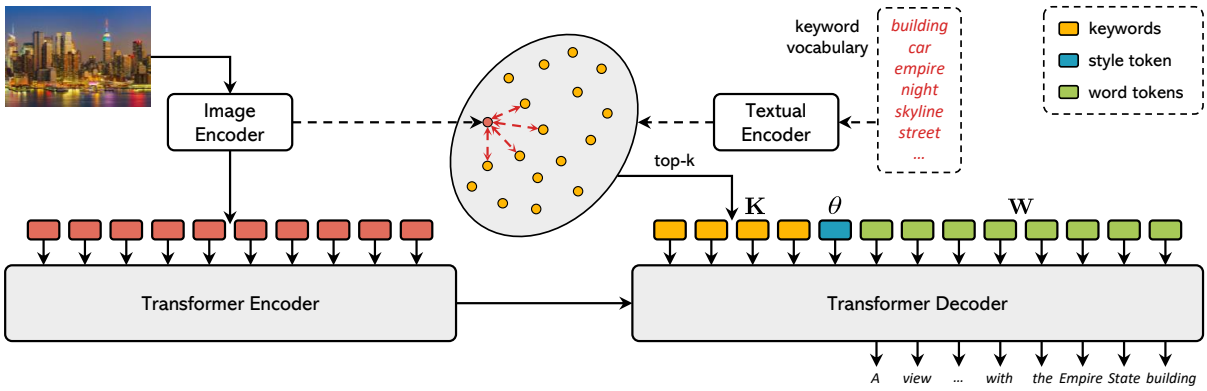


Fig. 3 Illustration of the overall structure of our approach, which is composed of an encoder module, a keyword-extraction module, and a decoder module.

their source and collection procedure. On average, though, they are more noisy and can comprise hashtags, comments and proper nouns.

Following this insight, we develop a model which is style-aware and can thus separate between the descriptive styles of the two aforementioned sets at generation time. Further, to enable the transfer of semantics learned from different sources, we employ textual keywords as a means to represent the content of an image regardless of its descriptive style. Formally, our approach considers a distribution

$$p(\boldsymbol{w}_t|\boldsymbol{w}_{\tau<t}, \mathbf{V}, \theta, \mathbf{K}), \tag{2}$$

where θ is a parameter encoding style and $\mathbf{K}$ a set of keywords encoding semantics.

3.1 Leveraging semantics and style

Extracting textual keywords through retrieval. Extracting a condensed textual representation of the visual input aims to promote an objective transfer between visual and textual features. Previous works have employed tags coming from an object detector (Anderson et al., 2018; P. Zhang et al., 2021), which however are limited in terms of the number of classes. Given the semantic breadth of web-scale datasets, it is instead crucial to scale beyond the limitation of object detection classes (Krishna et al., 2017). To this end, we cast the tagging problem as a cross-modal retrieval one, instead of a classification one.

Given a dictionary of keywords $\mathbf{Q}$, the set of predicted keywords for an input image $\mathbf{V}$ is obtained by selecting the k elements in $\mathbf{Q}$ with the highest similarity to $\mathbf{V}$, according to the matching function defined by a cross-modal retrieval model. Formally, being ϕ the similarity function defined by the cross-modal model, the set of keywords $\mathbf{K}$ is computed as

$$\mathbf{K} = \underset{\boldsymbol{q}\in\mathbf{Q}}{\text{argtop-}k}\ \phi(\mathbf{V}, \boldsymbol{q}), \tag{3}$$

where argtop-$k_{x\in\mathcal{S}} f(x)$ returns the elements in $\mathcal{S}$ which produce the k largest values in $\{f(x), x \in \mathcal{S}\}$.

The keywords dictionary $\mathbf{Q}$ must be large enough to ensure sufficient coverage with respect to the semantic distribution of web-collected datasets. To this aim, we construct $\mathbf{Q}$ by mining around 11.5k unigrams from COCO, Visual Genome, and the OpenWebText corpus, a public clone of OpenAI's WebText dataset (Radford et al., 2019). For computing cross-modal similarities, we leverage the multi-modal embedding space of CLIP (Radford et al., 2021), which can scale well in terms of the number of concepts. Embeddings for the keywords in $\mathbf{Q}$ are obtained through the CLIP text model and can be precomputed in advance. The embedding of an image is instead obtained through the CLIP image model and acts as a query of a kNN search. Although CLIP's encoder was trained on web-collected sentences, we found it to work well enough also in our case, in which it is fed with only one unigram. For efficiency reasons, we optimize this process

using an index for approximate search (Johnson, Douze, & Jégou, 2019). Compared to the tags extracted from object regions (Anderson et al., 2018; P. Zhang et al., 2021), the keywords we extract do not refer to local regions of the input image, but rather correspond to the image as a whole and have an increased semantic coverage with respect to object detectors trained on standard datasets.

Style token. To aid the generation process and separate content from style, we give the model awareness of the kind of dataset to which every training caption belongs. This is done with a "style token", which is implemented through a learnable token embedding and which can be concatenated to the representation of the keywords. For simplicity, we employ a style token with two possible values, one for encoding human-annotated sources (θ = `human-annotated`) and the second for encoding web-collected sources (θ = `web-collected`). Notice that, while θ = `human-annotated` actually refers to a uniform annotation style, θ = `web-collected` can be thought as containing a collection of heterogeneous descriptive styles, as web-collected sources might have different styles (*e.g.*, news captions and captions extracted from social media sites).

3.2 Architecture

Our approach represents each training image-caption pair as a quadruple $(\mathbf{V}, \mathbf{W}, \mathbf{K}, \theta)$ of image, ground-truth caption, keywords, and style token, where $\mathbf{V}$ is encoded with a set of fixed-length visual descriptors. The text input, including the caption and keywords, are tokenized lower-cased Byte Pair Encoding (BPE) (Sennrich, Haddow, & Birch, 2016).

For multimodal fusion, we define an encoder-decoder Transformer architecture (Vaswani et al., 2017) where each layer of the encoder comprises multi-head self-attention (MSA) and feed-forward layers, and each layer of the decoder includes multi-head self-attention (MSA), multi-head cross-attention (MSCA), and feed-forward layers. To enable text generation, the decoder employs sequence-to-sequence attention masks in each self-attention layer. The visual descriptors $\mathbf{V} = \{\boldsymbol{v}_i\}_{i=1}^N$ are encoded via bi-directional attention in the encoder, while textual keyword tokens $\mathbf{K} = \{\boldsymbol{k}_i\}_{i=1}^M$, token embeddings of the caption $\mathbf{W} = \{\boldsymbol{w}_i\}_{i=1}^L$, and the style token θ are inputs of the decoder, where N, M, and L indicates the number of visual embeddings, keywords, and caption tokens respectively. The overall network operates according to the following schema:

$$
\begin{aligned}
\text{encoder} \quad & \tilde{\boldsymbol{v}}_i = \mathrm{MSA}(\boldsymbol{v}_i, \mathbf{V}), \quad \tilde{\mathbf{V}} = \{\tilde{\mathbf{v}}_i\}_{i=1}^N \\
& \mathbf{O}_{\boldsymbol{k}_i} = \mathrm{MSCA}(\boldsymbol{k}_i, \tilde{\mathbf{V}}, \mathbf{K}) \\
\text{decoder} \quad & \mathbf{O}_{\theta} = \mathrm{MSCA}(\theta, \tilde{\mathbf{V}}, \mathbf{K} \cup \theta) \\
& \mathbf{O}_{\boldsymbol{w}_i} = \mathrm{MSCA}(\boldsymbol{w}_i, \tilde{\mathbf{V}}, \mathbf{K} \cup \theta \cup \{\boldsymbol{w}_t\}_{t=1}^i),
\end{aligned}
$$

where $\mathrm{MSA}(\boldsymbol{x}, \mathbf{Y})$ indicates a self-attention with $\boldsymbol{x}$ mapped to query and $\mathbf{Y}$ mapped to key-values, and $\mathrm{MSCA}(\boldsymbol{x}, \mathbf{Y}, \mathbf{Z})$ is a self-attention with $\boldsymbol{x}$ as query and $\mathbf{Z}$ as key-values, followed by a cross-attention with $\boldsymbol{x}$ as query and $\mathbf{Y}$ as key-values. $\mathbf{O}$ indicates the network output and $\cup$ indicates concatenation. We omit feed-forward layers and the dependency between consecutive layers for ease of notation.

Unlike a traditional decoder, the network is only trained to predict a left-shifted version of the caption tokens $\mathbf{W}$, while the sequence $\mathbf{K} \cup \theta$ is treated as a prompt. Different from prompting in pre-trained language models (Gao, Fisch, & Chen, 2021; Radford et al., 2019), this prompting strategy is explicitly employed while training the network. Further, in contrast to previous V&L pre-training works which adopted a bidirectional Masked Language Modeling objective that tends to be suboptimal for sequence generation, we train our network by following a unidirectional language modeling loss based on cross-entropy, *i.e.*

$$
\mathcal{L} = -\mathbb{E}_{(\mathbf{V},\mathbf{W})\sim\mathcal{D}} \left(\sum_{t=1}^{L} \log p(\mathbf{O}_{\boldsymbol{w}_t} | \mathbf{V}, \mathbf{K}, \theta, \boldsymbol{w}_{\tau<t}) \right), \tag{4}
$$

where $\mathcal{D}$ indicates the training dataset.

In the training stage, human-annotated and web-collected image-text pairs are fed through the model, each with its corresponding style token. To allow content transfer between the two types of data sources, it is important to maintain a sufficient balance between the two kind of data sources. To this end, we randomly select samples in each mini-batch to have at least 10% of image-caption pairs with θ = `human-annotated`. In our preliminary experiments, we verified that such a

Table 1 Statistics on the training corpus.

Source	Type	# Images	# Words	# Words in 0.1% Tail	Length (mean ± std)
COCO	Human annotated	112k	26,815	5,947	10.50 ± 2.42
Flickr30k	Human annotated	29k	17,798	1,793	12.34 ± 5.21
Open Images	Generated (COCO)	1.7M	7,693	4,050	10.09 ± 1.59
SBU	Flickr desc.	875k	222,393	10,053	12.20 ± 6.10
WIT	Wikipedia desc.	3.1M	905,095	281,32	9.21 ± 8.49
CC3M	Alt-texts	3.1M	47,422	21,100	9.58 ± 4.30
CC12M	Alt-texts	12.2M	450,594	189,792	18.28 ± 13.59
YFCC100M	Alt-texts	14.6M	2,384,078	383,942	26.31 ± 70.48
Overall		**35.7M**	**3,180,785**	**666,519**	**19.40 ± 47.11**

percentage is enough to enable a smooth content transfer between different data sources.

Inference. Once the model is trained, predictions are conditioned on the style token θ, which can be chosen according to the desired generation style (*i.e.* that of human-annotated captions or that of noisy web-collected ones). Given keywords and style token, at each time step t the model samples a token $\hat{\boldsymbol{w}}_t$ from the output probability distribution. This is then concatenated to previously predicted tokens to form a sequence $\{\hat{\boldsymbol{w}}_\tau\}_{\tau=1}^{t}$ which is employed as the input for the next iteration. Since the representation of output tokens does not depend on subsequent tokens, the past intermediate representations are kept in memory to avoid repeated computation and increase efficiency at prediction time.

Visual features. To obtain the set of visual features $\mathbf{V}$ for an image, we employ the same visual CLIP model employed for keyword retrieval (Radford et al., 2021). Compared to using features extracted from object detectors (Hu et al., 2022; P. Zhang et al., 2021), this strategy is beneficial both in terms of computational efficiency and feature quality. Specifically, we use a ViT-based visual encoder. In the original CLIP model, activations from the last layer of the encoder are discarded, except for those generated by the first query of the input sequence which are used to form a global descriptor. While global image vectors coming from CLIP have been used in concurrent captioning works (Mokady et al., 2021), we instead employ the entire grid of features coming from the layer, so to preserve spatial awareness and better feature granularity.

4 Experimental Evaluation

We conduct extensive experiments to validate the architectural choices behind our model and compare its performances with state-of-the-art solutions for image captioning.

4.1 Training sources

We train on a mixture of datasets with image-caption pairs, which are heterogeneous in terms of style and semantics, for a total of 35.7M images. Our mixture contains COCO (T.-Y. Lin et al., 2014), Flickr30k (Young et al., 2014), SBU (Ordonez et al., 2011), Conceptual Captions 3M (Sharma et al., 2018) and 12M (Changpinyo et al., 2021), WIT (Srinivasan et al., 2021), a subset of YFCC100M (Thomee et al., 2016), and a subset of Open Images (Kuznetsova et al., 2020).

The SBU dataset contains captions automatically collected from the Flickr website, while Conceptual Captions 3M and 12M have been obtained by cleaning image alt-text pairs from the web. The WIT dataset, instead, contains images extracted from Wikipedia together with alt texts. After filtering out all non-English pages, the dataset contains about 3.1M pairs. Finally, we use the subset of YFCC100M (Thomee et al., 2016) containing image descriptions (around 14.6M pairs), and 1.7M images from Open Images (Kuznetsova et al., 2020), automatically annotated with captions generated from OSCAR, following (P. Zhang et al., 2021).

In Table 1, we report detailed statistics on the data sources employed during training. Overall, the mixture used for training our model has three key features: (*i*) differently from the datasets employed in concurrent works (X. Li et al., 2020; P. Zhang et al., 2021), it contains only data for the image-captioning task, thus neglecting the use of data from ancillary tasks like VQA or GQA; (*ii*) it is made of publicly available data, thus allowing reproducibility, and does not employ proprietary data (Jia et al., 2021; Wang et al., 2022); (*iii*) overall, it contains 35.7 million images and around

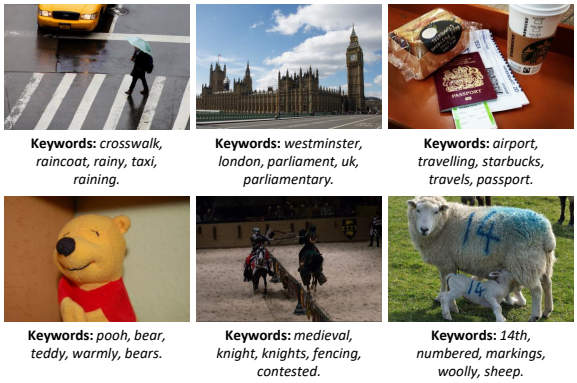


Fig. 4 Sample textual keywords extracted on COCO images.

0.6M long-tail words (*i.e.* that lie in the 0.1% of the distribution tail), making it sufficiently large and diverse to perform web-scale analyses on image captioning.

During training, we employ the style token for human-annotated sources (θ = `human-annotated`) when dealing with samples from COCO, Flickr30k, and Open Images, and the one for web-collected sources (θ = `web-collected`) when dealing with samples from SBU, WIT, CC3M, CC12M, and YFCC100M.

4.2 Implementation details

Architectural details. We devise three model configurations, varying the number of decoding layers L, model dimensionality d, and number of attention heads H: Tiny ($L = 3$, $d = 384$, $H = 6$, 52M params), Small ($L = 6$, $d = 512$, $H = 8$, 87M params), and Base ($L = 12$, $d = 768$, $H = 12$, 213M params). For all models, we employ CLIP-ViT-L/14 as visual feature and keyword extractor, three layers in the visual encoder, and five textual keywords. To represent words, we use lower-cased Byte Pair Encoding (BPE) (Sennrich et al., 2016) with a 49,152 vocab size and linearly project them to the input dimensionality of the model d. We employ classic sinusoidal positional encodings (Vaswani et al., 2017) to represent word positions. For efficiency, the maximum sequence length of the decoder is capped at 80.

Following recent literature (Cornia et al., 2020), we enrich all layers of our encoder with memory slots. Specifically, we extend the set of keys and values of each self-attention operation with 40 additional learnable vectors, which are independent of the input sequence and can encode a priori information retrieved through attention. For fair comparison, this also applies to all baselines and ablation studies presented in the following.

Training details. Training is performed using the LAMB optimizer (You et al., 2020) and following the learning rate scheduling strategy of (Vaswani et al., 2017) with a warmup equal to 6,000 iterations and multiplying the resulting learning rate by a factor of 5. We use a minibatch size of 1,080 and employ ZeRo memory offloading (Rajbhandari, Rasley, Ruwase, & He, 2020) and mixed-precision (Micikevicius et al., 2018). After training with cross-entropy, we also fine-tune some of our models on COCO using Reinforcement Learning. During this fine-tuning stage, we employ the SCST variant proposed in (Cornia et al., 2020) that sets the baseline reward equal to the mean of rewards of generated captions inside a beam. In this phase, we use the Adam optimizer (Kingma & Ba, 2015), a batch size equal to 80, and a fixed learning rate of 5×10^{-6}.

Keyword extraction details and visualization. As previously mentioned, the keyword dictionary is composed by extracting around 11.5k unigrams from COCO, Visual Genome, and the OpenWebText corpus[1]. During pre-processing, all unigrams are converted into lowercase, and proper names that identify persons are removed. Fig. 4 reports keywords extracted from sample images of the COCO dataset. As it can be seen, they provide significant and high-level information on the global content of the image. Compared to keywords extracted from object detectors, CLIP-based keywords tend to include more long-tail concepts.

Weight initialization. We initialize all weights by drawing inspiration from GPT-2 (Radford et al., 2019). All linear and embedding weights are initialized according to a uniform distribution and using the approach proposed by (Glorot & Bengio, 2010). Layer normalization weights are initialized to a constant value of 1. All biases are initialized to

[1] https://skylion007.github.io/OpenWebTextCorpus

Table 2 Comparison with the state of the art on the COCO test split in a single model setting.

	Fine-tuning							
	TF	SCST	**Training Images**	B-4	M	R	C	S
BLIP^{base} (J. Li et al., 2022)	✓	-	129M	39.7	-	-	133.3	-
BLIP^{large} (J. Li et al., 2022)	✓	-	129M	40.4	-	-	136.7	-
SimVLM^{base} (Wang et al., 2022)	✓	-	1.8B	39.0	32.9	-	134.8	24.0
SimVLM^{large} (Wang et al., 2022)	✓	-	1.8B	40.3	33.4	-	142.6	24.7
SimVLM^{huge} (Wang et al., 2022)	✓	-	1.8B	40.6	**33.7**	-	143.3	25.4
LEMON^{base} (Hu et al., 2022)	✓	✓	200M	41.6	31.0	-	142.7	25.1
LEMON^{large} (Hu et al., 2022)	✓	✓	200M	42.3	31.2	-	144.3	25.3
LEMON^{huge} (Hu et al., 2022)	✓	✓	200M	42.6	31.4	-	145.5	**25.5**
Ours^{tiny} (θ = human-annotated)	-	✓	35.7M	42.8	31.0	61.2	148.4	24.6
Ours^{small} (θ = human-annotated)	-	✓	35.7M	42.5	31.2	61.3	148.6	25.0
Ours^{base} (θ = human-annotated)	-	✓	35.7M	**42.9**	31.4	**61.5**	**149.6**	25.0
OSCAR^{base} (X. Li et al., 2020)	✓	✓	4.1M	40.5	29.7	-	137.6	22.8
OSCAR^{large} (X. Li et al., 2020)	✓	✓	4.1M	41.7	30.6	-	140.0	24.5
VinVL^{base} (P. Zhang et al., 2021)	✓	✓	5.8M	40.9	30.9	-	140.6	25.1
VinVL^{large} (P. Zhang et al., 2021)	✓	✓	5.8M	41.0	31.1	-	140.9	25.2
Ours^{tiny} (θ = human-annotated)	-	✓	5.8M (VinVL data)	42.9	31.1	61.3	147.1	24.9
Ours^{small} (θ = human-annotated)	-	✓	5.8M (VinVL data)	42.7	31.3	61.3	147.5	25.2
Ours^{base} (θ = human-annotated)	-	✓	5.8M (VinVL data)	**43.2**	**31.4**	**61.7**	**147.8**	**25.4**

0. We also employ the "Special Scaled Initialization" used in GPT-2[2] when initializing the output linear projection of each Transformer layer. Again, this also applies to all baselines.

4.3 Captioning performance

We firstly assess the performance of our model on human-annotated datasets, comparing with recent models trained using both web-collected and human-annotated data, *i.e.* OSCAR (X. Li et al., 2020), VinVL (P. Zhang et al., 2021), SimVLM (Wang et al., 2022), BLIP (J. Li et al., 2022), and LEMON (Hu et al., 2022)[3]. In contrast to these approaches that employ a teacher-forcing (TF) fine-tuning on COCO when comparing on COCO and nocaps, we only employ the RL fine-tuning (SCST), which we observed being lighter in terms of forgetting impact and avoids forgetting concepts learned on web-collected data. During generation, in this setting, we set the style token θ equal to the token embedding employed for human-annotated sources.

Evaluation is reported in terms of the classical captioning metrics: BLEU (Papineni, Roukos, Ward, & Zhu, 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (C.-Y. Lin, 2004), CIDEr (Vedantam, Lawrence Zitnick, & Parikh, 2015), and SPICE (Anderson, Fernando, Johnson, & Gould, 2016).

Performance on COCO. The performances of our approach on COCO are reported in the upper portion of Table 2, in a single model setting. As presented, the proposed method exhibits better performances than the compared methods, without requiring a teacher-forcing fine-tuning phase, using less training data and fewer parameters. The Tiny version of the proposed approach, for instance, overcomes the performance of SimVLM, BLIP, and LEMON in their Base, Large, and Huge configurations according to the BLEU-4, ROUGE, and CIDEr metrics. Increasing model size further augment the performances, up to 149.6 CIDEr points achieved by the Base version.

As additional comparison, we also report the performance of our models when trained with a dataset size that is directly comparable to that used for OSCAR and VinVL, which also employ the SCST fine-tuning stage. In particular, we employ the same datasets used for training VinVL, excluding VQA sources. Results are reported in the bottom part of Table 2, and show increased captioning metrics with respect to the compared approaches.

Results on nocaps. We then evaluate the capabilities of our model on the nocaps

[2] A reference implementation can be found in https://github.com/huggingface/transformers/blob/main/src/transformers/models/gpt2/modeling_gpt2.py#L493

[3] The number of parameters of these models is as follows: VinVL^{base} (135M), VinVL^{large} (370M), LEMON^{large} (338M), LEMON^{huge} (675M), BLIP^{base} (224M), BLIP^{large} (446M), SimVLM^{base} (86M), SimVLM^{large} (307M), SimVLM^{huge} (632M).

Table 3 Results on the nocaps dataset.

	Fine-tuning			Validation Set							
				in		near		out		overall	
	TF	SCST	Training Ims	C	S	C	S	C	S	C	S
OSCAR (X. Li et al., 2020)	✓	✓	112k (COCO)	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4
VIVO (Hu et al., 2020)	✓	✓	112k (COCO)	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4
VinVL (P. Zhang et al., 2021)	✓	✓	112k (COCO)	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1
BLIP^{base} (J. Li et al., 2022)	✓	-	129M	111.8	14.9	108.6	14.8	111.5	14.2	109.6	14.7
BLIP^{large} (J. Li et al., 2022)	✓	-	129M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8
SimVLM^{huge} (Wang et al., 2022)	✓	-	1.8B	113.7	-	110.9	-	115.2	-	112.2	-
LEMON^{large} (Hu et al., 2022)	✓	-	200M	116.9	**15.8**	113.3	15.1	111.3	14.0	113.4	15.0
LEMON^{huge} (Hu et al., 2022)	✓	-	200M	118.0	15.4	116.3	15.1	120.2	**14.5**	117.3	15.0
Ours^{tiny} (θ = human-annotated)	-	✓	35.7M	122.3	14.8	115.3	14.6	116.1	13.6	116.5	14.5
Ours^{small} (θ = human-annotated)	-	✓	35.7M	123.7	15.0	118.5	15.0	116.2	13.8	118.8	14.8
Ours^{base} (θ = human-annotated)	-	✓	35.7M	**124.8**	15.3	**119.6**	**15.2**	**120.3**	14.4	**120.5**	**15.1**
VinVL^{base} (P. Zhang et al., 2021)	✓	✓	5.8M	112.4	14.7	104.2	14.3	93.1	12.7	103.1	14.1
VinVL^{large} (P. Zhang et al., 2021)	✓	✓	5.8M	115.3	15.2	105.6	14.7	96.1	13.0	105.1	14.4
Ours^{tiny} (θ = human-annotated)	-	✓	5.8M (VinVL data)	121.4	14.9	115.7	14.8	110.6	13.5	115.5	14.6
Ours^{small} (θ = human-annotated)	-	✓	5.8M (VinVL data)	120.0	15.4	117.1	15.2	112.0	13.9	116.5	15.0
Ours^{base} (θ = human-annotated)	-	✓	5.8M (VinVL data)	**122.3**	**15.6**	**117.7**	**15.4**	**115.6**	**14.5**	**118.0**	**15.2**
				Test Set							
SimVLM^{huge} (Wang et al., 2022)	✓	-	1.8B	109.0	14.6	110.8	14.6	109.5	13.9	110.3	14.5
LEMON^{large} (Hu et al., 2022)	✓	-	200M	111.2	**15.6**	112.3	15.2	105.0	13.6	110.9	15.0
LEMON^{huge} (Hu et al., 2022)	✓	-	200M	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9
Ours^{tiny} (θ = human-annotated)	-	✓	35.7M	114.0	14.7	115.3	14.7	107.3	13.2	113.7	14.4
Ours^{small} (θ = human-annotated)	-	✓	35.7M	117.6	15.3	117.9	15.0	113.3	13.7	117.1	14.8
Ours^{base} (θ = human-annotated)	-	✓	35.7M	**118.8**	15.5	**120.4**	**15.4**	**114.0**	**14.1**	**119.1**	**15.2**
VinVL^{base} (P. Zhang et al., 2021)	✓	✓	5.8M	104.8	14.8	102.9	14.4	85.8	12.5	100.1	14.1
VinVL^{large} (P. Zhang et al., 2021)	✓	✓	5.8M	107.4	14.9	106.2	14.7	91.0	12.9	103.7	14.4
Ours^{tiny} (θ = human-annotated)	-	✓	5.8M (VinVL data)	115.2	15.2	115.2	15.0	106.3	13.8	113.6	14.8
Ours^{small} (θ = human-annotated)	-	✓	5.8M (VinVL data)	**117.2**	**15.8**	115.3	15.1	106.9	14.0	114.0	15.0
Ours^{base} (θ = human-annotated)	-	✓	5.8M (VinVL data)	116.0	15.6	**117.4**	**15.4**	**110.2**	**14.4**	**115.9**	**15.2**

dataset (Agrawal et al., 2019), which contains out-of-domain images with respect to COCO.

Table 3 reports the results for both the validation and test sets, at the top and bottom part respectively. As it can be seen, the proposed method achieves higher performances with respect to previous approaches when tested on nocaps, which confirms its capability of describing out-of-domain concepts. Also in this case, our approach tends to achieve better performance using fewer parameters than competitors. Our Tiny configuration, for instance, achieves higher generation quality than SimVLM^{huge} on both the validation and test sets, while the Small configuration is superior to LEMON^{huge}. It should be noted, nevertheless, that the performances on the out-of-domain portion of nocaps increase as the dimensionality of the model increases, from 116.1 to 120.3 CIDEr points when comparing the Tiny configuration to the Base configuration of the proposed approach. Overall, the Base version of our model overcomes all the reported competitors, testifying the advantage of the proposed strategies when describing out-of-domain concepts.

Fig. 5 Comparison of captions generated by VinVL and those generated by our approach on sample images from nocaps.

Table 4 Results on the CC3M validation split.

	TF Fine-tuning	Training Images	B-4	M	R	C	S
LEMON^{base} (Hu et al., 2022)	-	200M	10.1	11.9	-	108.1	19.8
LEMON^{base} (Hu et al., 2022)	✓	200M	10.1	12.0	-	111.9	20.5
LEMON^{large} (Hu et al., 2022)	✓	200M	10.8	12.3	-	117.4	21.0
LEMON^{huge} (Hu et al., 2022)	✓	200M	13.0	13.9	-	136.8	23.2
Ours^{tiny} (θ = `web-collected`)	✓	35.7M	10.6	13.1	30.0	121.3	23.0
Ours^{small} (θ = `web-collected`)	✓	35.7M	11.6	13.5	30.5	130.0	23.6
Ours^{base} (θ = `web-collected`)	-	35.7M	9.2	12.1	27.8	105.7	20.9
Ours^{base} (θ = `web-collected`)	✓	35.7M	**13.2**	**14.2**	**31.4**	**144.4**	**24.7**

Table 5 Ablation study on the COCO test split and nocaps validation split.

	Training Images: 35.7M							**Training Images: 5.8M (VinVL data)**						
	COCO					`nocaps`		**COCO**					`nocaps`	
	B-4	M	R	C	S	C	S	B-4	M	R	C	S	C	S
w/o web-collected data	40.6	30.0	59.9	139.4	23.9	88.9	12.6	40.6	30.0	59.9	139.4	23.9	88.9	12.6
w/o keywords and style token	42.5	30.6	60.8	145.8	24.2	108.8	13.6	42.4	30.9	61.0	145.4	24.8	108.2	14.2
w/o style token	42.5	30.7	60.9	146.5	24.3	110.1	13.6	42.5	31.0	61.1	146.2	24.8	108.6	14.2
Ours^{tiny} (θ = `human-annotated`)	**42.8**	**31.0**	**61.2**	**148.4**	**24.6**	**116.5**	**14.5**	**42.9**	**31.1**	**61.3**	**147.1**	**24.9**	**115.5**	**14.6**
w/o web-collected data	40.9	30.4	60.1	141.5	24.5	89.1	12.8	40.9	30.4	60.1	141.5	24.5	89.1	12.8
w/o keywords and style token	42.3	31.0	60.9	147.5	24.7	112.2	14.2	42.7	31.1	61.1	146.2	25.0	105.6	14.2
w/o style token	42.1	31.0	61.0	148.1	24.8	113.7	14.4	**43.0**	**31.3**	61.1	147.2	25.0	106.5	14.3
Ours^{small} (θ = `human-annotated`)	**42.5**	**31.2**	**61.3**	**148.6**	**25.0**	**118.8**	**14.8**	42.7	**31.3**	**61.3**	**147.5**	**25.2**	**116.5**	**15.0**
w/o web-collected data	41.4	30.2	60.2	142.0	24.1	89.2	12.6	41.4	30.2	60.2	142.0	24.1	89.2	12.6
w/o keywords and style token	42.6	31.3	61.2	147.9	**25.1**	114.7	14.8	42.6	31.1	61.2	146.8	24.8	109.4	14.3
w/o style token	42.5	**31.4**	61.2	149.2	25.0	116.6	14.9	42.7	31.3	61.5	147.0	25.2	109.5	14.5
Ours^{base} (θ = `human-annotated`)	**42.9**	**31.4**	**61.5**	**149.6**	25.0	**120.5**	**15.1**	**43.2**	**31.4**	**61.7**	**147.8**	**25.4**	**118.0**	**15.2**

We also investigate the behavior of the model when trained with the same data of VinVL, and thus on a smaller dataset. Compared to VinVL, which also employs SCST fine-tuning, our model is superior according to all metrics. Scaling the dataset size from 5.8M to 35.7M images brings, however, a significant improvement: from 147.8 to 149.6 CIDEr points on COCO, and from 118.0 to 120.5 CIDEr points on nocaps. To further validate the results on nocaps, we show in Fig. 5 sample images and corresponding textual descriptions generated by our model (with θ = `human-annotated`) in comparison to those generated by VinVL.

Results on CC3M. We also test our approach on Conceptual Captions 3M, which contains web-collected annotations. During generation, we set the style token θ to the token embedding employed for web-collected sources, so to have an appropriate generation style. We also test our model in a zero-shot setting (*i.e.* without any fine-tuning), and when fine-tuning with teacher forcing on this dataset. Results are reported in Table 4, in comparison with different configurations of the LEMON model. Without fine-tuning, our model achieves a better generation quality in terms of METEOR and SPICE, with slightly lower CIDEr scores. Table 4 also compares when running a fine-tuning stage, where our model overcomes LEMON according to all metrics, which again confirms the superiority of the proposed approach.

4.4 Ablation and analysis

Separation of semantics and style. In Table 5 we investigate the role of using retrieval-based keywords and the style token. Specifically, we report the results for all three versions of our model using different amounts of image-caption pairs during training. We notice that adding keywords alone provides an improvement on both COCO and nocaps and that the benefit is especially evident on out-of-domain captioning and with large models, confirming that retrieval-based keywords help to deal with out-of-domain concepts. The combination of keywords and the style token θ for human-annotated style, instead, further boosts performances on all model configurations. For instance, when using the Tiny configuration and the entire training dataset, it increases COCO performances from 145.8 CIDEr points to 148.4. The

Table 6 Comparison with different tags sources on the COCO test split and nocaps validation split. All models are trained on the COCO dataset only.

		COCO					nocaps	
	Tags/Keywords	B-4	M	R	C	S	C	S
Ours^{tiny}	None	40.6	30.0	59.9	139.4	23.9	88.9	12.6
Ours^{tiny}	Faster R-CNN	41.0	30.2	60.1	141.1	23.9	89.8	12.5
Ours^{tiny}	ResNeXt-152 C4	40.8	30.3	60.1	140.3	24.0	89.9	**12.7**
Ours^{tiny}	Retrieval-based	**41.1**	**30.4**	**60.2**	**141.9**	**24.1**	**90.7**	**12.7**
Ours^{base}	None	41.4	30.2	60.2	142.0	24.1	89.2	12.6
Ours^{base}	Faster R-CNN	40.5	30.3	60.3	140.5	23.7	90.2	13.3
Ours^{base}	ResNeXt-152 C4	40.5	30.2	60.0	141.7	23.8	88.4	12.3
Ours^{base}	Retrieval-based	**41.6**	**30.6**	**60.7**	**143.4**	**24.2**	**90.5**	**13.4**

Table 7 Results when conditioning on different styles on the COCO test split, the CC3M and WIT validation splits, and sample images from LAION-400M. Models are trained with cross-entropy loss on the entire training corpus.

		COCO			CC3M			WIT			LAION-400M		
	Style token	B-4	C	S	B-4	C	S	B-4	C	S	B-4	C	S
Ours^{base}	θ = `web-collected`	25.7	86.3	15.5	**9.2**	**105.7**	**20.9**	**2.0**	**25.9**	**6.5**	**4.1**	**58.1**	**9.4**
Ours^{base}	θ = `human-annotated`	**39.7**	**132.3**	**23.2**	5.0	58.8	14.6	0.9	12.3	4.2	2.1	31.2	6.9

same improvement is maintained when increasing model size, *e.g.* the Base configuration moves from 147.9 to 149.6 CIDEr points. Also in this case, the improvement is more significant on nocaps, where the Base configuration is improved from 114.7 to 120.5 CIDEr points, confirming that the two proposed strategies, together, help to transfer semantic concepts across descriptive styles. Similar improvements can also be noticed when training on a smaller dataset (*i.e.* on the same data of VinVL).

Comparison with other keywords strategies. We compare the proposed retrieval-based textual keywords with existing alternatives: tags extracted from Faster R-CNN trained on Visual Genome (Krishna et al., 2017) and tags extracted from a ResNeXt-152 C4, trained on the same data used in VinVL. Results are reported in Table 6 using both Tiny and Base versions, in which we train separate models with different keywords on COCO only, thus in a purely limited in-domain setting where transfer from web sources is not allowed. This choice is motivated by computational requirements, as the cost of running object detectors on large-scale data would have been intractable. As semantic transfer from web sources is not allowed, this setting is also less favorable for the proposed retrieval-based strategy. Employing the proposed tag approach, however, improves caption quality on both COCO and nocaps according to all metrics, bringing an improvement on the COCO dataset of around 0.8 and 1.4 CIDEr points respectively for the Tiny and Base versions.

Fig. 6 Sample images from the COCO dataset and captions predicted by the Base version of our model when conditioned on different styles.

Effect of the style token. We also assess the significance of the style token value. To this aim, we extract captions generated by our method when fed with both style token values. For this experiment, we report the results on the standard COCO test set and the CC3M and WIT validation splits. We also include the results on

Table 8 Performances on the FlickrStyle10k and SentiCap test splits.

	θ = humorous				θ = romantic				θ = positive				θ = negative			
	B-1	B-3	M	C	B-1	B-3	M	C	B-1	B-3	M	C	B-1	B-3	M	C
MSCap (Guo et al., 2019)	16.3	1.9	5.3	15.2	17.0	2.0	5.4	10.1	46.9	16.2	16.8	55.3	45.5	15.4	16.2	51.6
MemCap (Zhao et al., 2020)	19.9	4.3	7.4	19.4	21.2	4.8	8.4	22.4	50.8	17.1	16.6	54.4	48.7	19.6	15.8	60.6
SAN (G. Li, Zhai, Lin, & Zhang, 2021)	29.5	9.9	12.5	47.2	30.9	10.9	13.0	53.3	53.0	**23.4**	18.1	72.0	51.2	20.5	**17.6**	67.0
Ours^{tiny}	**32.9**	**12.9**	**15.6**	**59.9**	**34.4**	**14.3**	**15.3**	**61.6**	**54.5**	21.6	**19.1**	**91.7**	**51.4**	**22.3**	17.5	**83.9**

Table 9 User study results on sample images from the Open Images dataset. We report the percentage of times our full model (*i.e.* **Ours^{base}** with θ = `human-annotated`) is preferred against a competitor. Notably, our model is always preferred more than 50% of the times.

	Richness	**Coherency**	**Fluency**
vs. VinVL^{large} (P. Zhang et al., 2021)	59.1	58.8	53.2
vs. **Ours^{base}** (θ = `human-annotated`) – w/o web-collected data	70.7	71.3	58.0
vs. **Ours^{base}** (θ = `human-annotated`) – w/o keywords and style token	62.7	68.1	54.8
vs. **Ours^{base}** (θ = `web-collected`)	53.0	63.6	67.2

30k images randomly extracted from the LAION-400M dataset (Schuhmann et al., 2021) to validate the effectiveness of the style token on a famous web-collected dataset not employed during the training of our model. Results are reported in Table 7. As it can be seen, choosing the proper style token value significantly increases the quality of the generation on all considered datasets, highlighting that the model has learned to mimic both clean annotations and web-collected ones. As a complement to this experiment, we show in Fig. 6 some qualitative results on sample images from the COCO dataset when varying the style token during inference.

To further analyze the effectiveness of our model in correctly following a given conditioning signal in the form of style tokens, we experiment with the FlickrStyle10k (Gan et al., 2017) and SentiCap (Mathews et al., 2016) datasets which are commonly used in stylized captioning literature. While the former contains image-caption pairs with humorous and romantic styles, the latter contains images and corresponding captions with positive and negative sentiments. For fair comparison with other methods, we train the Tiny version of our model following the training protocol described in (G. Li et al., 2021) and without employing web-scale data. In this case, we use a different style token for each of the four styles contained in the two datasets plus one for factual captions (*i.e.* captions without a specific style or sentiment). Results are shown in Table 8 comparing our model with other previous captioning approaches focused on the generation of stylized captions. Notably, our approach can also handle a larger number of styles and perform competitively compared to methods specifically designed for the stylized captioning task.

User study. In addition to the standard quantitative metrics used in the previous analyses, we conduct a user study to fully validate the proposed model. In particular, we recruited 30 different participants and asked them to select either our full model or one of the competitors or baselines, judging in terms of (1) *richness of semantics*, (2) *coherency with the image*, and (3) *linguistic fluency*. Participants could also state that captions were equivalent on one or more evaluation axes. In this case, 0.5 points are given to both competitors.

We perform the user study on a random subset of Open Images (Kuznetsova et al., 2020) composed of 1,000 images. As shown in Table 9, adding large-scale data, keywords, and the style token increases the richness of the generated captions while maintaining coherency and fluency. Moreover, the use of a cleaned style (*i.e.* θ = `human-annotated`) also increases fluency and coherency with the input image. Even when comparing our model with the Large version of VinVL (P. Zhang et al., 2021), the user study results confirm the effectiveness of our approach and the improvement over the competitor especially in terms of richness of semantics and coherency with the image.

Table 10 Zero-shot performances on the VizWiz test split and the TextCaps validation split.

			VizWiz					TextCaps				
	Zero-Shot	**Training Images**	B-4	M	R	C	S	B-4	M	R	C	S
Up-Down (Anderson et al., 2018)	✓	112k	11.3	12.6	35.8	18.9	5.8	12.4	13.3	33.7	24.2	8.7
AoANet (Huang et al., 2019)	✓	112k	13.2	13.4	37.6	19.4	6.2	18.1	17.7	41.4	32.3	11.2
Up-Down (Anderson et al., 2018)	✗	23k/22k	19.8	18.4	43.2	49.7	12.2	20.1	17.8	**42.9**	41.9	11.7
AoANet (Huang et al., 2019)	✗	23k/22k	23.2	**20.3**	47.1	60.5	14.0	20.4	18.9	**42.9**	42.7	13.2
VinVL^{base} (P. Zhang et al., 2021)	✓	5.8M	16.9	15.8	41.1	34.7	9.9	17.3	16.5	38.9	41.2	13.1
VinVL^{large} (P. Zhang et al., 2021)	✓	5.8M	17.4	16.3	41.7	37.7	10.3	17.5	16.6	38.9	41.9	13.1
Ours^{tiny} (θ = human-annotated)	✓	5.8M (VinVL data)	22.5	18.7	45.7	56.9	14.1	20.5	17.9	41.0	53.0	14.6
Ours^{small} (θ = human-annotated)	✓	5.8M (VinVL data)	22.2	18.9	45.5	58.1	14.3	20.7	18.1	41.1	54.4	14.7
Ours^{base} (θ = human-annotated)	✓	5.8M (VinVL data)	22.5	19.2	45.9	59.6	14.9	20.6	18.2	41.2	55.4	15.0
Ours^{tiny} (θ = human-annotated)	✓	35.7M	23.6	19.3	46.4	65.6	14.8	20.7	18.0	41.1	58.6	14.6
Ours^{small} (θ = human-annotated)	✓	35.7M	24.5	19.9	47.2	70.2	15.3	21.9	18.9	42.3	66.0	15.4
Ours^{base} (θ = human-annotated)	✓	35.7M	**25.7**	**20.3**	**47.9**	**76.2**	**16.2**	**23.6**	**19.4**	42.8	**69.9**	**15.9**

Table 11 Long-tail description performances on the validation splits of Open Images, ImageNet-21K and CC3M.

		Open Images			
	Training Images	Long-tail Words	Named Words	CLIP-S	PAC-S
VinVL^{base} (P. Zhang et al., 2021)	5.8M	149	57	0.729	0.822
VinVL^{large} (P. Zhang et al., 2021)	5.8M	186	68	0.736	0.828
Ours^{base} (θ = human-annotated)	5.8M (VinVL data)	628	176	**0.761**	0.849
Ours^{base} (θ = human-annotated)	35.7M	**884**	**254**	0.759	**0.851**
		ImageNet-21K			
	Training Images	Long-tail Words	Named Words	CLIP-S	PAC-S
VinVL^{base} (P. Zhang et al., 2021)	5.8M	149	64	0.713	0.813
VinVL^{large} (P. Zhang et al., 2021)	5.8M	194	72	0.721	0.820
Ours^{base} (θ = human-annotated)	5.8M (VinVL data)	789	172	0.756	0.850
Ours^{base} (θ = human-annotated)	35.7M	**1152**	**261**	**0.759**	**0.851**
		CC3M			
	Training Images	Long-tail Words	Named Words	CLIP-S	PAC-S
VinVL^{base} (P. Zhang et al., 2021)	5.8M	84	46	0.732	0.824
VinVL^{large} (P. Zhang et al., 2021)	5.8M	95	45	0.735	0.829
Ours^{base} (θ = human-annotated)	5.8M (VinVL data)	572	112	0.777	**0.856**
Ours^{base} (θ = human-annotated)	35.7M	**581**	**162**	**0.778**	**0.856**

4.5 Zero-shot and long-tail description

Two of the main benefits of employing web-collected data are zero-shot generalization and the description of long-tail entities. In the following, we consider the VizWiz dataset (Gurari et al., 2020), which contains images originating from blind people, and TextCaps (Sidorov et al., 2020), with images containing text for zero-shot generalization. Both of them represent distinct visual and semantic distributions from the COCO ones. Further, we also investigate the capabilities of the proposed approach to deal with long-tail concepts and generate named entities.

Zero-shot performances on VizWiz and TextCaps. Table 10 shows a comparison when using Up-Down (Anderson et al., 2018) and AoANet (Huang et al., 2019) in a zero-shot manner trained on COCO and when fine-tuning them on the aforementioned datasets. We also compare with the Base and Large configurations of VinVL (P. Zhang et al., 2021). As shown, the proposed approach consistently outperforms the performances of Up-Down and AoANet when evaluated in a zero-shot setting. Interestingly, it also overcomes these approaches when trained on VizWiz and TextCaps, confirming that the model is capable of properly transferring semantic concepts learned from web-collected annotations. Further, our approach also beats VinVL in both configurations by a significant margin, highlighting the appropriateness of the proposed strategies

with respect to previous literature. For completeness, we also report the zero-shot performances of our approach when training on 5.8M images. Even when using less training data our approach still showcases good zero-shot prediction capabilities.

Long-tail and named entities description. We assess the capability of our approach to name long-tail concepts and named entities. In particular, we consider the validation sets of three datasets with a large variety of concepts: Open Images V6 (Kuznetsova et al., 2020) (subset with bounding boxes), ImageNet-21K (Ridnik, Ben-Baruch, Noy, & Zelnik-Manor, 2021), and CC3M. We count the number of unique words which do not appear in COCO at least 5 times (*i.e.* termed as long-tail words), and the number of named entities extracted using the spaCy NLP toolkit[4]. In this setting, we evaluate caption quality using the CLIP-Score (Hessel, Holtzman, Forbes, Bras, & Choi, 2021) and PAC-Score (Sarto, Barraco, Cornia, Baraldi, & Cucchiara, 2023) metrics, that both are based on CLIP features coming from a ViT-B/32 model, do not require ground-truth captions, and have a high correlation with human judgments. Table 11 shows the results of both versions of our Base model trained on 5.8M and 35.7M images, in comparison with VinVL. Our approach is capable of naming significantly more words that are outside of COCO and also consistently generates more named entities than VinVL. This improvement is further confirmed by the results in terms of CLIP-S and PAC-S which are significantly better than those obtained by VinVL. Reducing the amount of web-collected sources impacts performances, with a significant reduction in the number of long-tail words and named entities produced during generation. We notice, though, that even when using the same dataset size our approach beats VinVL according to all evaluation metrics in both its Base and Large configurations, further confirming the appropriateness of the proposed approach, and of avoiding teacher-forcing fine-tuning.

4.6 Qualitative results

Finally, we showcase the capabilities of our approach of generating pertinent captions with named entities (*i.e.* recognizable objects, places, and people) through some qualitative examples. In Fig. 1, on the second page, we compare with the baseline Tiny model trained without web-scale data and keywords (reported in Table 5). We observe that our approach correctly recognizes and names famous people, places, and trademarks like the *Burj Al Arab*, *Marilyn Monroe*, or the *Facebook logo*. The same can be observed in Fig. 5 and 7, where we compare with VinVL^{large} (P. Zhang et al., 2021) on images taken from nocaps, OpenImages, and CC3M. Again, our model can recognize and name long-tail concepts better than previous approaches, also recognizing famous people and places as for example the *Duke of Cambridge* and the *Pyramid of Djoser*. Additional qualitative results are reported in the Appendix.

Fig. 7 Comparison of captions generated by VinVL and those generated by our approach on sample images from CC3M and Open Images.

5 Conclusion

We proposed an approach for captioning images with fluent and pertinent descriptions while training on non-uniform data sources. Our approach relies on the separation of semantics and style and the usage of retrieval-based textual keywords, and allows to learn from noisy web-collected sources while maintaining a fluent descriptive style. Experimentally, it achieves state-of-the-art results on different datasets, including COCO, CC3M and nocaps, demonstrating its effectiveness in both in-domain and out-of-domain image captioning.

[4] https://spacy.io/

Acknowledgments. We thank CINECA for providing computational resources. This work has been supported by the PNRR-M4C2 project (PE00000013) "FAIR - Future Artificial Intelligence Research" funded by the European Commission and the PRIN "CREATIVE: CRossmodal understanding and gEnerATIon of Visual and tExtual content" co-funded by the Italian Ministry of University and Research (CUP B87G22000460001).

References

Agrawal, H., Desai, K., Chen, X., Jain, R., Batra, D., Parikh, D., ... Anderson, P. (2019). nocaps: novel object captioning at scale. *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... others (2022). Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems.*

Anderson, P., Fernando, B., Johnson, M., Gould, S. (2016). SPICE: Semantic Propositional Image Caption Evaluation. *Proceedings of the European Conference on Computer Vision.*

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops.*

Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R. (2022). The Unreasonable Effectiveness of CLIP Features for Image Captioning: An Experimental Analysis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

Barraco, M., Stefanini, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R. (2022). CaMEL: Mean Teacher Learning for Image Captioning. *Proceedings of the International Conference on Pattern Recognition.*

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems.*

Changpinyo, S., Sharma, P., Ding, N., Soricut, R. (2021). Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Chen, T., Zhang, Z., You, Q., Fang, C., Wang, Z., Jin, H., Luo, J. (2018). "Factual"or"Emotional": Stylized Image Captioning with Adaptive Learning and Attention. *Proceedings of the European Conference on Computer Vision.*

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., ... Liu, J. (2020). UNITER: Learning UNiversal Image-TExt Representations. *Proceedings of the European Conference on Computer Vision.*

Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An Image is

Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations.*

Gan, C., Gan, Z., He, X., Gao, J., Deng, L. (2017). StyleNet: Generating Attractive Visual Captions With Styles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Gao, T., Fisch, A., Chen, D. (2021). Making Pre-trained Language Models Better Few-shot Learners. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the international conference on artificial intelligence and statistics.*

Guo, L., Liu, J., Yao, P., Li, J., Lu, H. (2019). MSCap: Multi-Style Image Captioning with Unpaired Stylized Text. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N. (2020). Captioning Images Taken by People Who Are Blind. *Proceedings of the European Conference on Computer Vision.*

Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y. (2021). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L. (2022). Scaling Up Vision-Language Pre-training for Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Hu, X., Yin, X., Lin, K., Wang, L., Zhang, L., Gao, J., Liu, Z. (2020). VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning. *Proceedings of the AAAI Conference on Artificial Intelligence.*

Huang, L., Wang, W., Chen, J., Wei, X.-Y. (2019). Attention on Attention for Image Captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., ... Duerig, T. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *Proceedings of the International Conference on Machine Learning.*

Johnson, J., Douze, M., Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Trans. on Big Data*, *7*(3), 535–547,

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Kim, W., Son, B., Kim, I. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *Proceedings of the International Conference on Machine Learning.*

Kingma, D.P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Proceedings of the International Conference on Learning Representations.*

Klein, F., Mahajan, S., Roth, S. (2021). Diverse Image Captioning with Grounded Style. *Proceeding of the DAGM German Conference on Pattern Recognition.*

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Fei-Fei, L. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, *123*(1), 32–73,

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ... Ferrari, V. (2020). The Open Images Dataset V4.

International Journal of Computer Vision, *128*(7), 1956–1981,

Li, G., Zhai, Y., Lin, Z., Zhang, Y. (2021). Similar Scenes arouse Similar Emotions: Parallel Data Augmentation for Stylized Image Captioning. *Proceedings of the ACM International Conference on Multimedia.*

Li, J., Li, D., Savarese, S., Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *Proceedings of the International Conference on Machine Learning.*

Li, J., Li, D., Xiong, C., Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *Proceedings of the International Conference on Machine Learning.*

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... others (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *Proceedings of the European Conference on Computer Vision.*

Li, Y., Pan, Y., Yao, T., Mei, T. (2022). Comprehending and Ordering Semantics for Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshops.*

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. *Proceedings of the European Conference on Computer Vision.*

Lu, J., Batra, D., Parikh, D., Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural Information Processing Systems.*

Mathews, A., Xie, L., He, X. (2016). SentiCap: Generating Image Descriptions with Sentiments. *Proceedings of the AAAI Conference on Artificial Intelligence.*

Mathews, A., Xie, L., He, X. (2018). Semstyle: Learning to generate stylised image captions using unaligned text. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... Wu, H. (2018). Mixed Precision Training. *Proceedings of the International Conference on Learning Representations.*

Mokady, R., Hertz, A., Bermano, A.H. (2021). ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*, ,

Ordonez, V., Kulkarni, G., Berg, T. (2011). Im2Text: Describing Images Using 1 Million Captioned Photographs. *Advances in Neural Information Processing Systems.*

Pan, Y., Yao, T., Li, Y., Mei, T. (2020). X-Linear Attention Networks for Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the International Conference on Machine Learning.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.*

Rajbhandari, S., Rasley, J., Ruwase, O., He, Y. (2020). ZeRO: Memory optimizations

Toward Training Trillion Parameter Models. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.*

Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V. (2017). Self-Critical Sequence Training for Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L. (2021). ImageNet-21K Pretraining for the Masses. *Advances in Neural Information Processing Systems.*

Sarto, S., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R. (2023). Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Sarto, S., Cornia, M., Baraldi, L., Cucchiara, R. (2022). Retrieval-Augmented Transformer for Image Captioning. *Proceedings of the International Conference on Content-Based Multimedia Indexing.*

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems.*

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., ... Komatsuzaki, A. (2021). LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *Advances in Neural Information Processing Systems.*

Sennrich, R., Haddow, B., Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Sharma, P., Ding, N., Goodman, S., Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., ... Keutzer, K. (2022). How Much Can CLIP Benefit Vision-and-Language Tasks? *Proceedings of the International Conference on Learning Representations.*

Sidorov, O., Hu, R., Rohrbach, M., Singh, A. (2020). TextCaps: A Dataset for Image Captioning with Reading Comprehension. *Proceedings of the European Conference on Computer Vision.*

Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M. (2021). WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. *ACM SIGIR Conference on Research and Development in Information Retrieval.*

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R. (2022). From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 539–559,

Tan, H., & Bansal, M. (2019). LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., ... Li, L.-J. (2016). YFCC100M: The new data in multimedia research. *Communications of the ACM*, *59*(2), 64–73,

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems.*

Vedantam, R., Lawrence Zitnick, C., Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y. (2022). SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. *Proceedings of the International Conference on Learning Representations.*

Xu, H., Yan, M., Li, C., Bi, B., Huang, S., Xiao, W., Huang, F. (2021). E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning.*

Yan, M., Xu, H., Li, C., Bi, B., Tian, J., Gui, M., Wang, W. (2021). Grid-VLP: Revisiting Grid Features for Vision-Language Pre-training. *arXiv preprint arXiv:2108.09479*, ,

Yang, X., Tang, K., Zhang, H., Cai, J. (2019). Auto-Encoding Scene Graphs for Image Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., ... Hsieh, C.-J. (2020). Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. *Proceedings of the International Conference on Learning Representations.*

Young, P., Lai, A., Hodosh, M., Hockenmaier, J. (2014). 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,

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y. (2022). CoCa: Contrastive Captioners are Image-Text Foundation Models. *arXiv preprint arXiv:2205.01917*, ,

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., ... Gao, J. (2021). VinVL: Revisiting visual representations in vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., ... others (2022). OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, ,

Zhao, W., Wu, X., Zhang, X. (2020). MemCap:Memorizing Style Knowledge for Image Captioning. *Proceedings of the AAAI Conference on Artificial Intelligence.*

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J. (2020). Unified Vision-Language Pre-Training for Image Captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence.*

Appendix A Additional Qualitative Results

We report different qualitative results obtained on images from nocaps (Fig. A1), VizWiz (Fig. A2), TextCaps (Fig. A3), CC3M and Open Images (Fig. A4). We observe how our model can describe objects, people, and scenes with a significantly increased level of detail when compared to the current state of the art and regardless of the dataset. Also, our approach qualitatively appears to be less prone to hallucination and can constantly generate fluent textual descriptions.

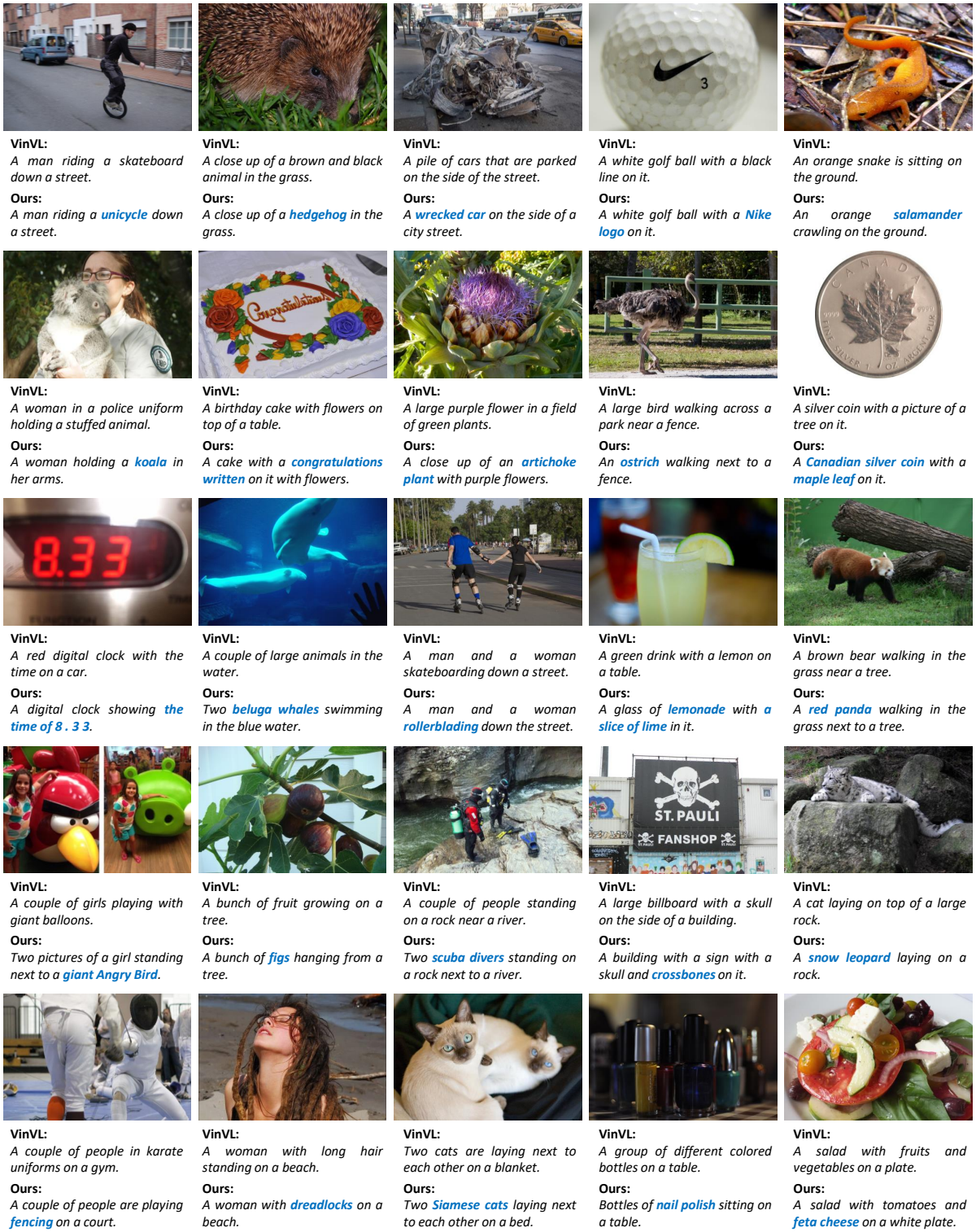
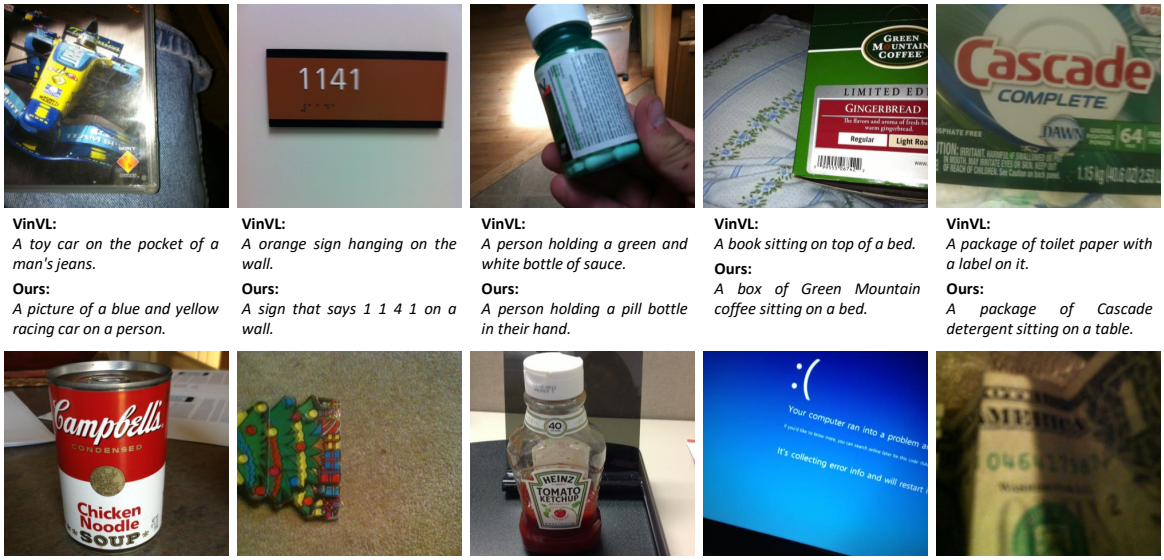


Fig. A1 Sample descriptions generated on nocaps images.



VinVL: *A toy car on the pocket of a man's jeans.*
Ours: *A picture of a blue and yellow racing car on a person.*

VinVL: *A orange sign hanging on the wall.*
Ours: *A sign that says 1 1 4 1 on a wall.*

VinVL: *A person holding a green and white bottle of sauce.*
Ours: *A person holding a pill bottle in their hand.*

VinVL: *A book sitting on top of a bed.*
Ours: *A box of Green Mountain coffee sitting on a bed.*

VinVL: *A package of toilet paper with a label on it.*
Ours: *A package of Cascade detergent sitting on a table.*

VinVL: *A can of coffee sitting on top of a table.*
Ours: *A can of chicken noodle soup sitting on a table.*

VinVL: *A colorful block of gummy bears on the floor.*
Ours: *A picture of a colorful Christmas tree on the floor.*

VinVL: *A bottle of cough syrup sitting on a counter.*
Ours: *A bottle of tomato ketchup sitting on a desk.*

VinVL: *A blue sky with a half moon and a line of light.*
Ours: *A computer screen with a sad face on it.*

VinVL: *A close up of a book with a building in the background.*
Ours: *A dollar bill sitting on top of a table.*

Fig. A2 Sample descriptions generated on images from the VizWiz dataset.

VinVL: *A picture of a movie poster on a wall.*
Ours: *A poster of a group of Lego figures with lightsaber.*

VinVL: *A red stop sign with graffiti on it.*
Ours: *A red stop sign with the word quest written on it.*

VinVL: *A poster on a wall in a room.*
Ours: *A no butts on the beach poster in a window.*

VinVL: *A busy city street with a lot of street signs.*
Ours: *A stop here on red signal sign on a city street.*

VinVL: *A red and white airplane is flying in the sky.*
Ours: *A Turkish airlines plane flying in the sky.*

VinVL: *A group of blue jerseys hanging on a shelf.*
Ours: *A group of Chelsea shirts hanging in a locker room.*

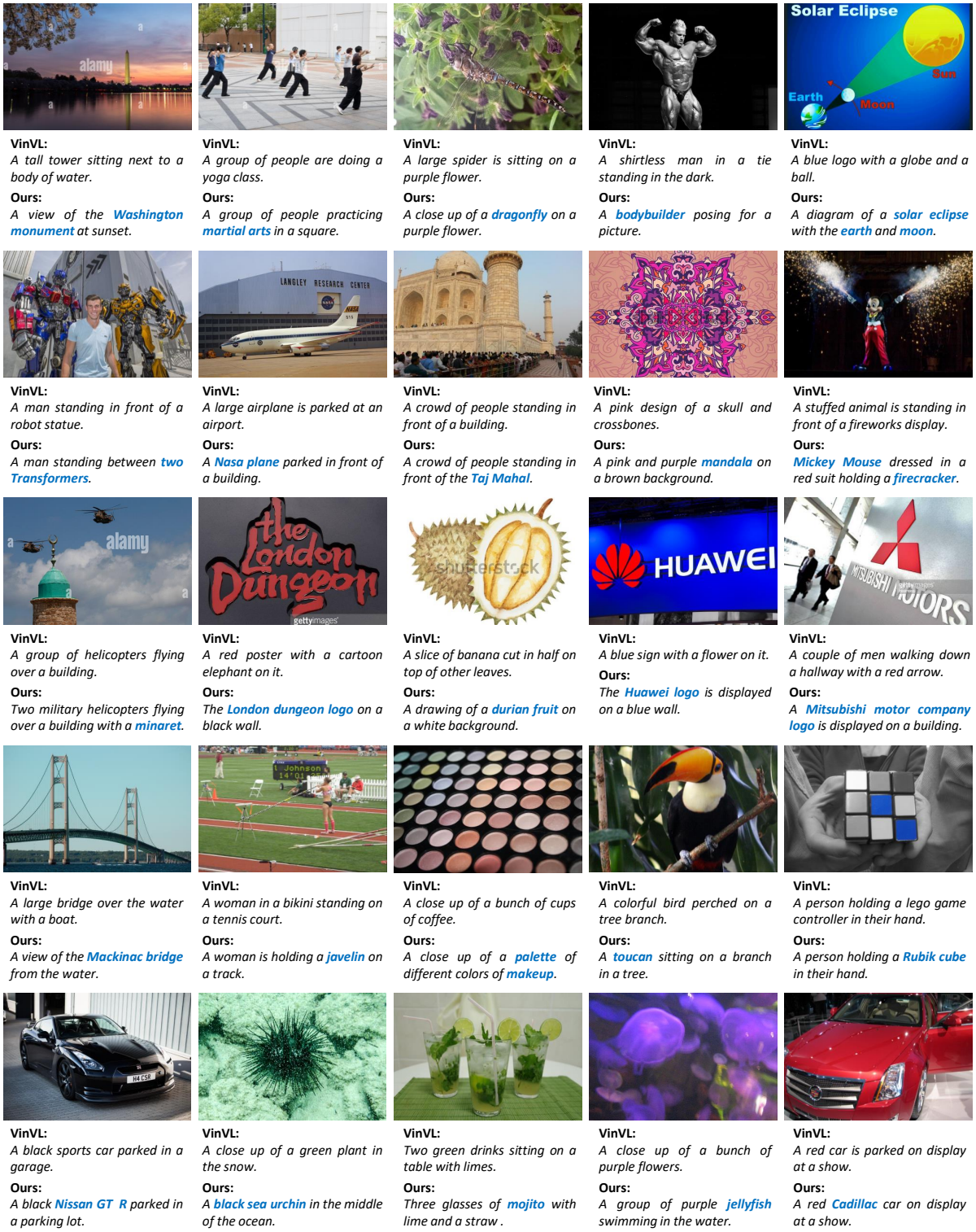
VinVL: *A book about the ghost of a man.*
Ours: *A cover of a Star Wars story with a man holding a sword.*

VinVL: *A large billboard with a crown on the side of a building.*
Ours: *A building with a sign that says Drupalcon Copenhagen.*

VinVL: *A couple of books sitting next to each other.*
Ours: *A blue book with the title half hours at the seaside.*

VinVL: *A motel sign on the side of a building.*
Ours: *A car wash shop with an American flag flying in front.*

Fig. A3 Sample descriptions generated on images from the TextCaps dataset.



VinVL: *A tall tower sitting next to a body of water.*
Ours: *A view of the **Washington monument** at sunset.*

VinVL: *A group of people are doing a yoga class.*
Ours: *A group of people practicing **martial arts** in a square.*

VinVL: *A large spider is sitting on a purple flower.*
Ours: *A close up of a **dragonfly** on a purple flower.*

VinVL: *A shirtless man in a tie standing in the dark.*
Ours: *A **bodybuilder** posing for a picture.*

VinVL: *A blue logo with a globe and a ball.*
Ours: *A diagram of a **solar eclipse** with the **earth** and **moon**.*

VinVL: *A man standing in front of a robot statue.*
Ours: *A man standing between **two Transformers**.*

VinVL: *A large airplane is parked at an airport.*
Ours: *A **Nasa plane** parked in front of a building.*

VinVL: *A crowd of people standing in front of a building.*
Ours: *A crowd of people standing in front of the **Taj Mahal**.*

VinVL: *A pink design of a skull and crossbones.*
Ours: *A pink and purple **mandala** on a brown background.*

VinVL: *A stuffed animal is standing in front of a fireworks display.*
Ours: *Mickey Mouse dressed in a red suit holding a **firecracker**.*

VinVL: *A group of helicopters flying over a building.*
Ours: *Two military helicopters flying over a building with a **minaret**.*

VinVL: *A red poster with a cartoon elephant on it.*
Ours: *The **London dungeon logo** on a black wall.*

VinVL: *A slice of banana cut in half on top of other leaves.*
Ours: *A drawing of a **durian fruit** on a white background.*

VinVL: *A blue sign with a flower on it.*
Ours: *The **Huawei logo** is displayed on a blue wall.*

VinVL: *A couple of men walking down a hallway with a red arrow.*
Ours: *A **Mitsubishi motor company logo** is displayed on a building.*

VinVL: *A large bridge over the water with a boat.*
Ours: *A view of the **Mackinac bridge** from the water.*

VinVL: *A woman in a bikini standing on a tennis court.*
Ours: *A woman is holding a **javelin** on a track.*

VinVL: *A close up of a bunch of cups of coffee.*
Ours: *A close up of a **palette** of different colors of **makeup**.*

VinVL: *A colorful bird perched on a tree branch.*
Ours: *A **toucan** sitting on a branch in a tree.*

VinVL: *A person holding a lego game controller in their hand.*
Ours: *A person holding a **Rubik cube** in their hand.*

VinVL: *A black sports car parked in a garage.*
Ours: *A black **Nissan GT R** parked in a parking lot.*

VinVL: *A close up of a green plant in the snow.*
Ours: *A **black sea urchin** in the middle of the ocean.*

VinVL: *Two green drinks sitting on a table with limes.*
Ours: *Three glasses of **mojito** with lime and a straw .*

VinVL: *A close up of a bunch of purple flowers.*
Ours: *A group of purple **jellyfish** swimming in the water.*

VinVL: *A red car is parked on display at a show.*
Ours: *A red **Cadillac** car on display at a show.*

Fig. A4 Sample descriptions generated on images from the CC3M and Open Images datasets.