Text-Only Training for Image Captioning using Noise-Injected CLIP

David Nukrai **Ron Mokady** **Amir Globerson**
Blavatnik School of Computer Science, Tel Aviv University

Abstract

We consider the task of image-captioning using only the CLIP model and additional text data at training time, and no additional captioned images. Our approach relies on the fact that CLIP is trained to make visual and textual embeddings similar. Therefore, we only need to learn how to translate CLIP textual embeddings back into text, and we can learn how to do this by learning a decoder for the frozen CLIP text encoder using only text. We argue that this intuition is "almost correct" because of a gap between the embedding spaces, and propose to rectify this via noise injection during training. We demonstrate the effectiveness of our approach by showing SOTA zero-shot image captioning across four benchmarks, including style transfer. Code, data, and models are available at https://github.com/DavidHuji/CapDec.

1 Introduction

Vision and language are closely intertwined, as they are two ways of describing the world. This raises the potential for developing models that map images and text into a shared semantic space. Indeed, this approach has recently achieved great success with models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). These models use *parallel* image-text data to train a joint representation, where the embeddings of image-text pairs are similar. Such models have been employed for various vision-language tasks.

Image captioning is a key task in vision-language perception. Yet, training image captioning models typically requires large datasets of captioned images, and these are challenging to collect. Furthermore, it is not clear how one could adapt a pretrained vision-language model to generate captions in new styles. In this work, we present an approach to captioning that only requires CLIP and text data, and generates styled captions using only unpaired textual examples from that style. This alleviates the need for paired text-image data, and also allows for simple style transfer.

A first approach one could consider for this setting is to train a decoder model to reconstruct texts from their respective CLIP embeddings, and at inference use this decoder to decode image embeddings. However, we observed that this approach fails at inference, and we conjecture this is due to the known domain gap between the image and text modalities (Liang et al., 2022). We propose a simple approach to mitigate this, by injecting noise into the embedding during training. This has the effect of creating a ball in embedding space that will map to the same caption, and corresponding image-embedding is more likely to be inside this ball, as illustrated in Fig. 1.a.

We evaluate our "Captioning via Decoding" (CapDec) method extensively, showing that it works well on several image captioning tasks, including standard, cross-domain, and style-guided captioning. Overall, our main contributions are as follows: 1) A simple and intuitive approach to learning a captioning model based on CLIP and additional text training data, but no images for training. 2) Evaluation of CapDec on image captioning tasks, including generating captions in various styles, shows it outperforms other methods which use the same supervision.

2 Related Work

Image captioning methods (Chen and Zitnick, 2014; Chen et al., 2017; Yang et al., 2019; Herdade et al., 2019; Luo et al., 2021; Tsimpoukelli et al., 2021) typically extract visual features using a pre-trained network. These are passed to a textual decoder that produces the final captions. To bridge the gap between vision and language, other works employ pre-training to create a shared latent space of vision and text (Tan and Bansal, 2019; Laina et al., 2019; Lu et al., 2019; Li et al., 2020; Zhou et al., 2020; Zhang et al., 2021; Wang et al., 2021;

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4055–4063
December 7-11, 2022 ©2022 Association for Computational Linguistics

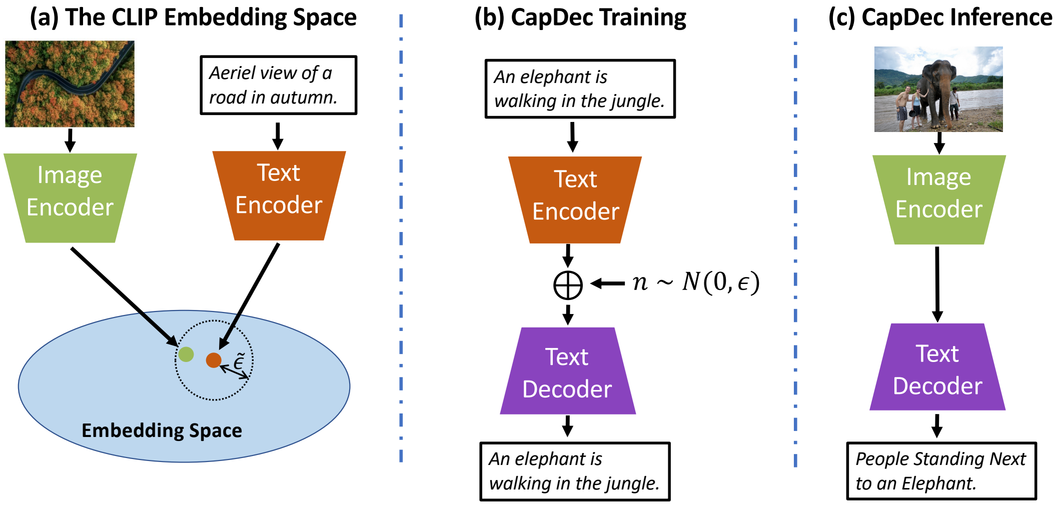

(a) The CLIP Embedding Space
Aeriel view of a road in autumn.
Image Encoder
Text Encoder
$\tilde{\epsilon}$
Embedding Space
(b) CapDec Training
An elephant is walking in the jungle.
Text Encoder
$n \sim N(0, \epsilon)$
Text Decoder
An elephant is walking in the jungle.
(c) CapDec Inference
Image Encoder
Text Decoder
People Standing Next to an Elephant.

Figure 1: **Overview of our CapDec captioning approach.** **(a)** An illustration of the CLIP joint embedding space. Embedded text is relatively close to its corresponding visual embedding, but with a certain gap. **(b)** CapDec trains a model that decodes the CLIP embedding of text T back to text T, after noise-injection. The encoders remain frozen. **(c)** At inference, CapDec simply decodes the embedding of an image using the trained decoder.

Hu et al., 2022). However, all of these approaches require extensive training and large paired datasets that are hard to collect. Gan et al. (2017) and Zhao et al. (2020) have suggested style-guided captioning, but also employ training over paired data.

CLIP (2021) marked a turning point in vision-language perception, and has been utilized for vision-related tasks by various distillation techniques (Gu et al., 2021; Song et al., 2022; Jin et al., 2021; Gal et al., 2021; Xu et al., 2021; Khandelwal et al., 2022). Recent captioning methods use CLIP for reducing training time (Mokady et al., 2021), improved captions (Shen et al., 2021; Luo et al., 2022a,b; Cornia et al., 2021; Kuo and Kira, 2022), and in zero-shot settings (Su et al., 2022; Tewel et al., 2022). However, zero-shot techniques often result in inferior performance, as the produced captions are not compatible with the desired target style, which is usually dictated by a dataset. In this work, we suggest a new setting, where we adapt CLIP to image captioning using only textual data. As a result, we can easily adapt captions to any desired caption style given instances of text in that style. Concurrent work by Su et al. (2022) efficiently produces high-quality captions with the minimal supervision of text-only pre-training by employing CLIP-induced score at inference. Our approach is arguably simpler and also outperforms Su et al. (2022) empirically. Note that Zhou et al. (2021) have also employed noise-injection, but for the opposite problem of CLIP-based text-free text-to-image generation.

3 Method

Text-Only Training. Our goal is to learn a model that produces a caption for a given image I. Unlike supervised approaches, we assume that during training we only have access to a set of texts $\mathcal{T}$. These can be obtained by harvesting a text corpus. We next introduce notation for the CLIP model. Given an image I let $\phi(I) \in \mathbb{R}^d$ be its embedding, and given a text T let $\psi(T) \in \mathbb{R}^d$ be its embedding. For converting a vector $\boldsymbol{v} \in \mathbb{R}^d$ into a caption, we use a textual decoder $C(\boldsymbol{v})$ consisting of a lightweight mapping network and a pretrained auto-regressive language model, as suggested in Mokady et al. (2021).

We train the decoder as follows (except for the noise-injection which we introduce below). Each text $T \in \mathcal{T}$ is first mapped to CLIP space via $\psi(T)$ and then decoded back into a text via $C(\psi(T))$. We would like this decoding to be similar to the original text T. Namely, our training objective is a reconstruction of the input text from CLIP's textual embedding. At inference, given an image I we simply apply the decoder to $\phi(I)$, returning the caption $C(\phi(I))$.

Noise-Injected CLIP Embeddings. We observed that the above training scheme results in inaccurate captions during inference. We conjecture this is because the embeddings of the text and image modalities are separated by a domain gap, as shown in Liang et al. (2022). As a result, while text reconstruction is successful during training,

(A) **Image Captioning**

Model	MS-COCO					Flickr30k				
	B@1	B@4	M	R-L	CIDEr	B@1	B@4	M	R-L	CIDEr
Fully Supervised Approaches										
BUTD	77.2	36.2	27.0	56.4	113.5	-	27.3	21.7	-	56.6
UniVLP	-	36.5	28.4	-	116.9	-	30.1	23.0	-	67.4
ClipCap	74.7	33.5	27.5	-	113.1	-	21.7	22.1	47.3	53.5
Oscar	-	36.5	30.3	-	123.7	-	-	-	-	-
LEMON	-	40.3	30.2	-	133.3	-	-	-	-	-
Weakly or Unsupervised Approaches										
ZeroCap	49.8	7.0	15.4	31.8	34.5	44.7	5.4	11.8	27.3	16.8
MAGIC	56.8	12.9	17.4	39.9	49.3	44.5	6.4	13.1	31.6	20.4
CapDec	**69.2**	**26.4**	**25.1**	**51.8**	**91.8**	**55.5**	**17.7**	**20.0**	**43.9**	**39.1**

(B) **Cross-Domain Captioning**

	Flickr30k $\Longrightarrow$ MS-COCO					MS-COCO $\Longrightarrow$ Flickr30k				
	B@1	B@4	M	R-L	CIDEr	B@1	B@4	M	R-L	CIDEr
MAGIC	41.4	5.2	12.5	30.7	18.3	46.4	6.2	12.2	31.3	17.5
CapDec	**43.3**	**9.2**	**16.3**	**36.7**	**27.3**	**60.2**	**17.3**	**18.6**	**42.7**	**35.7**

Table 1: **Results for image captioning. (A)** We use captions from the COCO and Flickr30k to train CapDec and evaluate on the datasets the captions were taken from. We report results for fully supervised methods that train on captioned images, and on methods that use no training text (ZeroCap), or just training text and no images (CapDec and MAGIC). **(B)** Similar setting to (A), but in cross-domain setup where training text is taken from one dataset, and evaluation is done on the second dataset.

inference fails when using image embeddings instead. If image-text pairs were available, we could attempt to learn a mapping between these domains. Nevertheless, as we aim for text-only training, we shall seek a different approach.

Specifically, we assume that the visual embedding corresponding to a text embedding lies somewhere within a ball of small radius ϵ around the text embedding (see Fig. 1). We would like all text embeddings in this ball to decode to the same caption, which should also correspond to the visual content mapped to this ball. We implement this intuition by adding zero-mean Gaussian noise of STD ϵ to the text embedding before decoding it.

The value of ϵ is calculated by estimating the spread of captions corresponding to the same image. Specifically, we set ϵ to the mean ℓ_∞ norm of embedding differences between five captions that correspond to the same image. We estimated this based on captions of only 15 MS-COCO images. Since this calculation requires very few captions and there is no need to recalculate it for every new dataset, we do not view it as additional supervision.

Our overall training objective is thus to minimize:

$$\sum_{T\in\mathcal{T}} \ell(C(\psi(T)+\boldsymbol{n}),T)\,, \qquad (1)$$

where $\boldsymbol{n} \in \mathbb{R}^d$ is a random standard Gaussian noise with standard-deviation ϵ and ℓ is an auto-regressive cross-entropy loss for all tokens in T. We train just the parameters of the textual decoder C, while the encoder $\psi()$ is kept frozen. The noise is sampled independently at each application of the encoder.

4 Results

We next evaluate CapDec on several captioning tasks, demonstrating state-of-the-art results. See supplementary for additional details.

Image Captioning. We compare CapDec caption quality to several baselines with different supervision levels, as presented in Tab. 1(A). Here, all methods were trained end evaluated over the same dataset, using the commonly used MS-COCO (Lin et al., 2014; Chen et al., 2015) and Flickr30k (Young et al., 2014). We begin by evaluating fully supervised techniques: BUTD (Anderson et al., 2018), UniVLP (Zhou et al., 2020), ClipCap (Mokady et al., 2021), Oscar (Li et al., 2020), and Lemon (Hu et al., 2022). As expected, these achieve a better score than CapDec, as they exploit the additional supervision of image-text pairs. Nevertheless, compared to the unsupervised ap-

Model	Romantic				Humorous			
	B@1	B@3	M	C	B@1	B@3	M	C
StyleNet	13.3	1.5	4.5	7.2	13.4	0.9	4.3	11.3
MemCap	21.2	4.8	8.4	22.4	19.9	4.3	7.4	19.4
CapDec + Image-Text Pre-training	27.9	8.9	12.6	52.2	29.4	8.8	13.2	55.1
CapDec + Text-Only Pre-training	23.0	4.6	9.1	27.4	22.7	4.3	9.7	29.0
CapDec	21.4	5.0	9.6	26.9	24.9	6.0	10.2	34.1

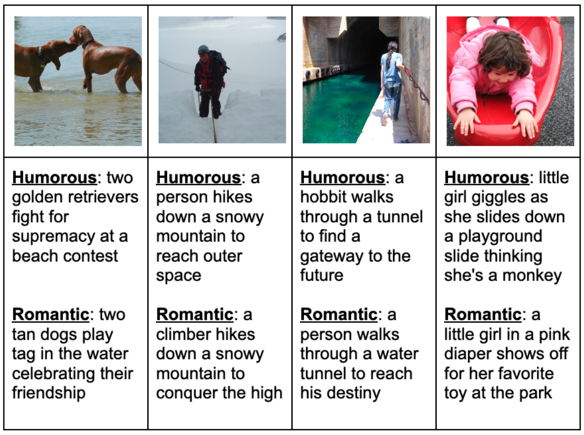
Table 2: **Style-Guided captioning results on FlickrStyle10K** (Gan et al., 2017).

proaches of MAGIC (Su et al., 2022) and ZeroCap (Tewel et al., 2022), CapDec achieves superior scores. Note that ZeroCap does not require any training data, while MAGIC requires text data similar to our setting.

Cross-Domain Captioning. We test our generalization ability by training on one dataset while evaluating on another, as in Su et al. (2022). Again, as can be seen in Tab. 1(B), CapDec outperforms MAGIC (Su et al., 2022), which uses the same supervision as CapDec.

Style-Guided Captioning. Several works (Zhao et al., 2020; Gan et al., 2017) have studied the task of adapting a captioning model to a new style, such as “romantic” or “humorous”. Since collecting paired examples for each style requires great effort, these consider the setting where the new style is only learned from text. This is easy to do in our setting, since we can train the decoder on any given style text. Fig. 2 shows captions generated with CapDec in several styles (same setting and data as in Zhao et al. (2020)). Tab. 2 reports quantitative results for this setting, showing CapDec outperforms other baselines. To further analyze our approach, we present our results without pre-training (i.e., training on styled data only), with a text-only pre-training over COCO, and with text-image pre-training over COCO (similar to (Zhao et al., 2020)). As can be seen, we outperform (Zhao et al., 2020) even with considerably less supervision at pre-training. Moreover, both other variations improve results, demonstrating that CapDec can effectively use additional training data where available.

The Effect of Noise Level. A key element of CapDec is noise injection before decoding. To demonstrate the effect of noise, we report results as a function of the noise variance ϵ^2 in Fig. 3. It can be seen that too little or too much noise is suboptimal. We note that the noise variance we chose, ϵ^2=0.016,[1] is based only on text, and not on the results in Fig. 3 which are shown for analysis purposes only.



Humorous: two golden retrievers fight for supremacy at a beach contest **Romantic**: two tan dogs play tag in the water celebrating their friendship	**Humorous**: a person hikes down a snowy mountain to reach outer space **Romantic**: a climber hikes down a snowy mountain to conquer the high	**Humorous**: a hobbit walks through a tunnel to find a gateway to the future **Romantic**: a person walks through a water tunnel to reach his destiny	**Humorous**: little girl giggles as she slides down a playground slide thinking she's a monkey **Romantic**: a little girl in a pink diaper shows off for her favorite toy at the park

Figure 2: Example for styled captions of CapDec on FlickrStyle10K (Gan et al., 2017).

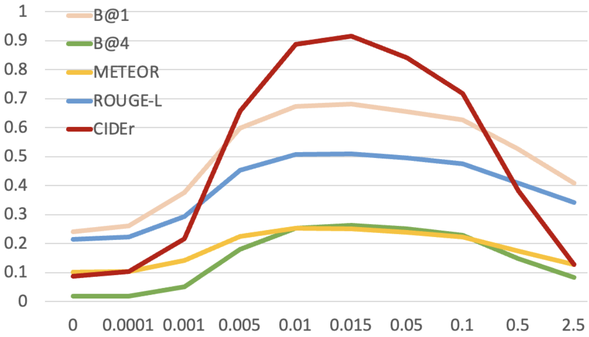


Figure 3: The effect of the noise variance on MS-COCO performance.

5 Noise Injection Analysis

Noise-injection is a well-known technique for improving generalization (Reed and MarksII, 1999; Bishop, 1995; An, 1996; Vincent et al., 2010), and can be viewed as a data augmentation mechanism

[1] As mentioned in Sec. 3, we estimated the optimal STD by the mean infinity-norm of embedding differences between captions that correspond to the same image, which is ϵ=$\sqrt{0.016}$

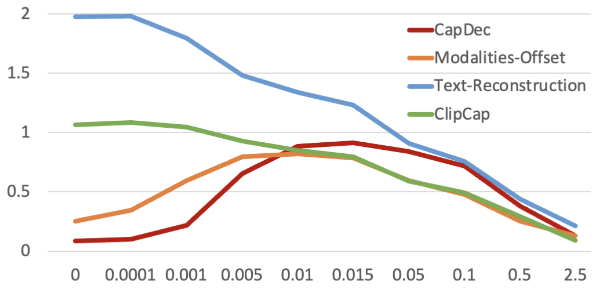


Figure 4: Analysis of performance of different methods as a function of the noise level (see Sec.5). We show the CiDER metric (higher is better), as other metrics show similar trends. CapDec here is the same as in Fig.3

(Goodfellow et al., 2016). In our case, the use of noise was also meant to address the modality-gap observed in Liang et al. (2022). In order to examine the specific effect of noise, we perform additional evaluations on COCO and show the results in Fig.4.

Text-Reconstruction: We encode COCO captions using CLIP text embedding and decode them using the learned CapDec model. This does not involve images at all, and is meant to test whether noise injection simply serves as regularization for text auto-encoding. Fig.4 shows that adding noise does not help, and thus suggests that noise is not merely functioning as augmentation.

ClipCap: Recall that ClipCap is trained on joint text-image pairs (Mokady et al., 2021). Here we trained ClipCap by adding noise to the image embeddings during training. It can be seen that noise does not improve performance, again suggesting that improvement is due to its specific role in domain-gap correction.

Modalities Offset: Given sufficient training paired-data, one could presumably learn the modalities-gap and correct for it. Here we test a simple approximation of the gap, that does not require image-text data to be paired, by calculating the shift between the mean of text embeddings and the mean of image embeddings in COCO. Then, given an image, we add the shift to its embedding to "correct" for this gap, and apply the CapDec trained decoder to the resulting embedding. Had this mapping been perfect, CapDec would not have needed additional noise injection. The results in Fig.4 show that the offset-correction does outperform CapDec at $\epsilon^2 < 0.01$, but underperforms overall. This suggests that the gap was not perfectly estimated, and that noise injection still serves to mitigate it. We leave it for future research to consider a more complex or fully-supervised model that learns the modality-gap explicitly.

6 Conclusion

The image captioning task has been extensively studied, with considerable progress in recent years. However, the number of available training datasets, containing image-text pairs, is still rather limited. Consequently, image captioning models inherit the limitations of their training data, such as biases (Hendricks et al., 2018) or confinement to neutral style (Gan et al., 2017). In this work, we suggest a new paradigm, where a generic vision-language model (e.g., CLIP) is adapted to image captioning using a text-only dataset. Furthermore, we demonstrate a simple and intuitive technique to overcome the inherent domain gap of CLIP (Liang et al., 2022). For future work, we plan to study text-only training for other tasks, such as visual question answering and visual scene graph generation.

7 Ethics Statement

Image captioning models are notorious for their internal biases (Hendricks et al., 2018). These biases are usually inherited from the training data itself. We observe that since balancing a text-only dataset is much more feasible than collecting balanced text-image pairs, CapDec can be used to mitigate those biases. For instance, consider the problem of a dataset containing significantly more images of snowboarding men than women. Collecting more images requires substantial effort while replacing "man" with "woman" (and their synonyms) in all captions is quite simple. Therefore, our text-only training might mitigate some of the inherited bias.

8 Limitations

We observe that although CapDec achieves superior results compared to the baselines that use only text at training, it is still outperformed by fully supervised baselines. Since CLIP captures rich semantics in its latent space, we believe that text-only training can be further improved up to the almost same quality as supervised techniques in future work. In addition, note that CapDec relies on CLIP and a language model both of which were pre-trained on large English corpora. Therefore, we find the important task of extending CapDec's capabilities to other languages to be a significant challenge.

Acknowledgments

This work was supported by the Blavatnik Interdisciplinary Research Center (ICRC). We thank to Amir Hertz for sharing relevant code parts from his work on ClipCap (Mokady et al., 2021).

References

Guozhong An. 1996. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Chris M Bishop. 1995. Training with noise is equivalent to Tikhonov regularization. *Neural computation*, 7(1):108–116.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667.

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

Xinlei Chen and C Lawrence Zitnick. 2014. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*.

Marcella Cornia, Lorenzo Baraldi, Giuseppe Fiameni, and Rita Cucchiara. 2021. Universal captioner: Long-tail vision-and-language model training through content-style separation.

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

Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. 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.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Ying Jin, Yinpeng Chen, Lijuan Wang, Jianfeng Wang, Pei Yu, Zicheng Liu, and Jenq-Neng Hwang. 2021. Is object detection necessary for human-object interaction recognition? *arXiv preprint arXiv:2107.13083*.

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

Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Chia-Wen Kuo and Zsolt Kira. 2022. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17979.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424.

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

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-level collaborative transformer for image captioning. *arXiv preprint arXiv:2101.06462*.

Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. 2022a. A frustratingly simple approach for end-to-end image captioning.

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

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Russell Reed and Robert J MarksII. 1999. *Neural smithing: supervised learning in feedforward artificial neural networks*. Mit Press.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12).

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le

Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. 2021. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2019. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4250–4260.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. 2020. Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12984–12992.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2021. Lafite: Towards language-free training for text-to-image generation. *arXiv preprint arXiv:2111.13792*.

A Appendix

A.1 Implementation Details

We use the RN-50x4 backbone for CLIP image encoder, and GPT-2 (large) as our language model (implementation of Wolf et al.(Wolf et al., 2020)). Following ClipCap (Mokady et al., 2021), for the decoder architecture, we use a transformer-based (Vaswani et al., 2017) mapping network where we set the CLIP embedding length of $K = 40$ with additional $K = 40$ constants tokens and use 8 multi-head self-attention layers with 8 heads each. For optimization, we employed AdamW (Kingma and Ba, 2015) with weight decay as introduced by Loshchilov et al. (Loshchilov and Hutter, 2017), with a learning rate of $2e^{-5}$ and 5000 warm-up steps.

A.2 Datasets and Evaluation Metrics

When evaluating over MS-COCO (Chen et al., 2015) and Flickr30k (Plummer et al., 2015), we followed Karpathy(Karpathy and Fei-Fei, 2015) split, similar to (Su et al., 2022) and (Mokady et al., 2021). Considering the FlickrStyle10K (Gan et al., 2017) dataset, we followed (Zhao et al., 2020), and split the dataset randomly to $6/7$, and $1/7$ of training and test sets, correspondingly. For qualitative evaluation, we employ the commonly used BLEU (Papineni et al., 2002) (B@1,B@4), METEOR (Denkowski and Lavie, 2014) (M), ROUGE-L (Lin and Och, 2004) (R-L), and CIDEr (Vedantam et al., 2015) (C) metrics.

A.3 Qualitative Comparison

All qualitative scores were reproduced or obtained from the works of (Su et al., 2022) and (Zhao et al., 2020) after carefully validating we use the same splits. Our metrics implementation is adapted from the official implementation of (Li et al., 2020).