Linear Alignment of Vision-language Models for Image Captioning

Fabian Paischer [1], Markus Hofmarcher [2], Sepp Hochreiter [1], Thomas Adler [1],
[1] ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning,
[2] JKU LIT SAL eSPML Lab, Institute for Machine Learning,
Johannes Kepler University, Linz, Austria
paischer@ml.jku.at

Abstract

Recently, vision-language models like CLIP have advanced the state of the art in a variety of multi-modal tasks including image captioning and caption evaluation. Many approaches adapt CLIP-style models to a downstream task by training a mapping network between CLIP and a language model. This is costly as it usually involves calculating gradients for large models. We propose a more efficient training protocol that fits a linear mapping between image and text embeddings of CLIP via a closed-form solution. This bypasses the need for gradient computation and results in a lightweight captioning method called ReCap, which can be trained up to 1000 times faster than existing lightweight methods. Moreover, we propose two new learning-based image-captioning metrics that build on CLIP score along with our linear mapping. We evaluate ReCap on MS-COCO, Flickr30k, VizWiz, and MSRVTT. ReCap achieves performance comparable to state-of-the-art lightweight methods on established metrics while outperforming them on our new metrics, which are better aligned with human judgement than established ones.

1. Introduction

Vision-language models (VLMs) are usually trained to align images and texts in a joint bi-modal embedding space. This enables their application to a variety of downstream tasks such as image-text retrieval [16], image captioning [13], few-shot classification [14], and caption evaluation [4]. As one of the most prominent VLMs, CLIP [15] has been pre-trained on a large-scale web dataset consisting of image-text pairs and advanced the state of the art across a variety of vision-language tasks. One of the most important downstream tasks is image captioning.

Adapting CLIP to a downstream task is generally costly in terms of both computational resources and data collection. In the context of image captioning, related works train mapping networks between CLIP and a generative language model (LM) [10, 12, 13, 16, 21]. Inspired by these recent successes, we aim at linearly aligning image and text embeddings of CLIP-style models to leverage them for retrieval augmentation in image captioning. This use case of CLIP is based on cross-modal retrieval via cosine similarity. Artetxe et al. [1] showed that a linear solution to a constrained least-squares problem is equivalent to maximizing the cosine similarity (under the same constraint). Leveraging this insight, we maximize the cosine similarity of image-text correspondences from the downstream dataset with respect to a constrained linear mapping. As this problem has a closed-form solution, we are able to align CLIP to the downstream data without the need for gradient computation. This makes our proposed method extremely versatile as training can be conducted within seconds on CPU.

We propose a fast and easily deployable method for adapting CLIP to a target domain. Given a set of image-text pairs representing a downstream task, we embed them in the joint embedding space of CLIP. To mitigate the modality gap [7] we re-align them by computing a linear mapping via a constrained least-squares solution (cf. Fig. 1, a). The linear mapping introduces only 0.0016% of trainable parameters compared to the original CLIP model. We demonstrate that this technique can be readily incorporated into an image captioning pipeline via retrieval augmentation (cf. Fig. 1, b). Given a new image, we embed it in the CLIP embedding space and apply our mapping to retrieve similar captions to it from a datastore filled with captions. These captions are then formatted to a prompt which is provided to a LM to generate a new caption for the image. We call the resulting method **Re**trieval-augmented **Cap**tioner (ReCap). Further, established caption evaluation metrics are rule-based and rely on heuristics, such as n-gram matching to reference captions. However, prior work has shown that pre-trained VLMs can act as reference-free caption evaluation metrics and correlate well with human judgement [4]. We build on this finding and propose two new learning-based image-captioning metrics that use our linear alignment to adapt CLIP-based metrics toward a downstream

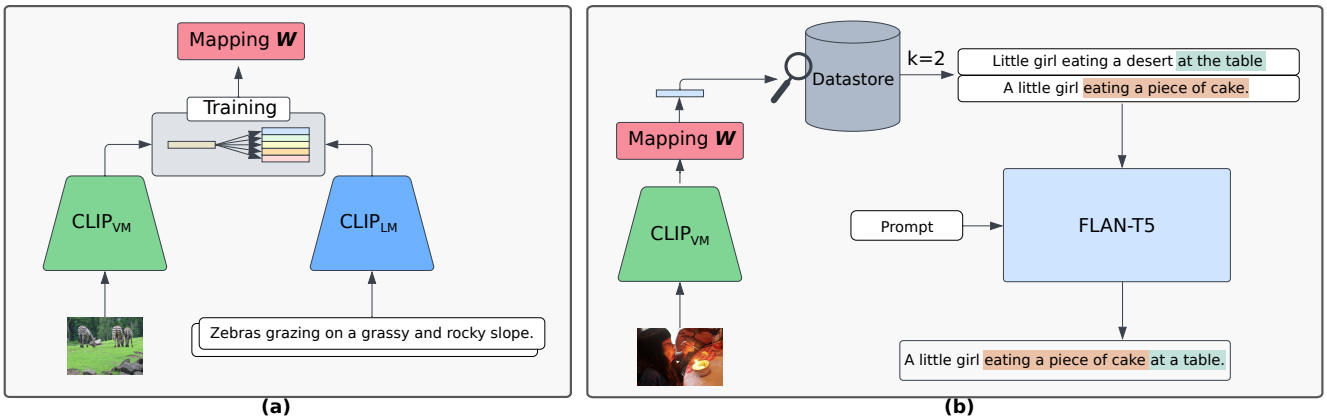

Figure 1. **(a)** We train a linear mapping $\boldsymbol{W}$ to align the image and text embeddings of CLIP toward a dataset. **(b)** On inference, we employ the mapping to retrieve captions from a datastore and provide these along with a prompt to a FLAN-T5 model to generate a new caption.

dataset. This consistently improves correlation with human judgement.

We evaluate ReCap on the MS-COCO [8], Flickr30k [20], VizWiz [3], and MSRVTT [19] datasets. By means of established metrics, ReCap achieves performance competitive to lightweight baselines that require over 1000 times more training effort on MS-COCO and Flickr30k, while outperforming the only available retrieval-augmented baseline, SmallCap [16], on VizWiz and MSRVTT. Further, we evaluate the correlation of our proposed metrics with human judgement on two datasets, Flickr8k-Expert and Flickr8k-Crowdflower [5]. Our metrics consistently improve over the CLIP-based metrics that rely on cosine similarity [4] and set a new state of the art in three out of four categories. By means of our newly proposed metrics, ReCap outperforms competitors on all four datasets.

2. Methods

We propose a linear alignment method for CLIP that optimizes cosine similarity between image-text pairs coming from a downstream dataset. The linear alignment computes a mapping in closed form under an orthogonality constraint. Therefore, it is very efficient to compute and easy to implement while only adding a relatively small set of trainable parameters. We elaborate on our linear alignment technique in more detail in Sec. 2.1. In Sec. 2.2 we introduce a lightweight image captioning pipeline based on our linear alignment without any further training. Sec. 2.3 introduces two new metrics, aCLIP-S, a reference-free metric, and RefaCLIP-S, a reference-based metric, both of which are based on the CLIP score [4] in combination with our linear alignment.

2.1. Linear Alignment of CLIP

Since our downstream use of CLIP involves retrieval via cosine similarity, we want to maximize the cosine similarity between image and text embeddings of a downstream dataset. To this end, we assume access to a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{c}_i)\}$ that provides image-text pairs, e.g., MS-COCO [8]. First, we embed the images of the training split $\mathcal{D}_{\text{Train}} \subset \mathcal{D}$ using a CLIP vision encoder $\phi : \mathcal{X} \to \mathbb{R}^d$, where $\mathcal{X}$ is the pixel space and d denotes the dimension of the joint CLIP embedding space. This results in an image embedding matrix $\boldsymbol{F}_{\mathcal{D}_{\text{Train}}} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top \in \mathbb{R}^{n \times d}$, where $\boldsymbol{f}_i = \phi(\boldsymbol{x}_i)$ for $i \in \{1, \ldots, n\}$ and $n = |\mathcal{D}_{\text{Train}}|$. Similarly, we embed the corresponding captions via the CLIP text encoder $\psi : \mathcal{T} \to \mathbb{R}^d$, where $\mathcal{T}$ is the space of tokenized strings, yielding a caption embedding matrix $\boldsymbol{E}_{\mathcal{D}_{\text{Train}}} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)^\top \in \mathbb{R}^{n \times d}$.

We employ a linear mapping $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ to re-align CLIP according to $\mathcal{D}_{\text{Train}}$. We aim to find a mapping $\boldsymbol{W}$ that projects an image embedding to the text embedding space such that its closest neighbor in terms of cosine similarity is its ground-truth caption. Yet, a closed-form solution for $\boldsymbol{W}$ to maximize the cosine similarity is unknown. By constraining $\boldsymbol{W}$ to be an orthogonal matrix, however, we obtain equivalence to the least-squares objective, that is

$$\begin{aligned} \boldsymbol{W}^* &= \underset{\boldsymbol{W} \text{ s.t. } \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}}{\arg\max} \sum_i \operatorname{cossim}(\boldsymbol{e}_i, \boldsymbol{W} \boldsymbol{f}_i) \\ &= \underset{\boldsymbol{W} \text{ s.t. } \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}}{\arg\min} \sum_i \|\boldsymbol{e}_i - \boldsymbol{W} \boldsymbol{f}_i\|_2^2 = \boldsymbol{V} \boldsymbol{U}^\top, \end{aligned}$$

where $\boldsymbol{V}$ and $\boldsymbol{U}$ are the orthogonal matrices of the singular value decomposition of $\boldsymbol{E}_{\mathcal{D}_{\text{Train}}}^\top \boldsymbol{F}_{\mathcal{D}_{\text{Train}}} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top$ and $\operatorname{cossim}(\cdot, \cdot)$ is the usual cosine similarity for vectors. This fact has been shown by Artetxe et al. [1] and we also provide a proof in the appendix for convenience. The solution to the constrained optimization problem is well known as *orthogonal procrustes* in the literature [17]. Notably, the size of $\boldsymbol{W}$ varies with the dimensionality d. Therefore, different CLIP encoders result in different amounts of parameters introduced by $\boldsymbol{W}$.

2.2. Retrieval-augmented Image Captioning

We utilize $\boldsymbol{W}$ for retrieval augmentation, where the retrieval datastore $\mathcal{C}$ contains captions of the training set $\mathcal{D}_{\text{Train}}$. Then we project a given image to the caption embedding space and retrieve its nearest neighbors. Given an image $\boldsymbol{x} \in \mathcal{X}$, we compute an embedding $\phi(\boldsymbol{x})$ and select the set $\mathcal{K}$ of top-k captions by

$$\mathcal{K} = \arg\max_{\boldsymbol{c}\in\mathcal{C}}^{k} \text{cossim}(\psi(\boldsymbol{c}), \boldsymbol{W}\phi(\boldsymbol{x})), \tag{1}$$

where $\arg\max^k$ denotes an extension of the $\arg\max$ operator returning the arguments of the k largest elements of a set. This way, we obtain a set of captions that provide a textual description of the image $\boldsymbol{x}$. We feed the retrieved captions $\mathcal{K}$ to a generative LM as context along with a prompt to generate a new caption for the image $\boldsymbol{x}$ (cf. Fig. 1, b). We use nucleus sampling [6] to obtain a set $\mathcal{S}$ of l candidate captions for the image $\boldsymbol{x}$ and select the candidate which yields the highest cosine similarity by

$$\arg\max_{\boldsymbol{s}\in\mathcal{S}} \text{cossim}(\psi(\boldsymbol{s}), \boldsymbol{W}\boldsymbol{f}). \tag{2}$$

The only trainable parameters of ReCap are $\boldsymbol{W}$ which only requires computing a closed-form solution on CPU. Specifically, computing $\boldsymbol{W}$ requires $\mathcal{O}(d^3)$ steps.

2.3. Image Caption Evaluation Metric

Given an image $\boldsymbol{x}$ and a candidate caption $\boldsymbol{c}$ we define the aligned CLIP score as

$$\text{aCLIP-S}(\boldsymbol{c}, \boldsymbol{x}) = \max\{\text{cossim}(\psi(\boldsymbol{c}), \boldsymbol{W}\phi(\boldsymbol{x})), 0\}. \tag{3}$$

Notably, aCLIP-S is reference-free, meaning it can be applied to any candidate without access to ground-truth human annotations, i.e., reference captions. In case a set $\mathcal{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \dots\}$ of reference captions is available, we can incorporate those into our score, which results in a reference-based metric

$$\text{RefaCLIP-S}(\boldsymbol{c}, \mathcal{R}, \boldsymbol{x}) = \tag{4}$$
$$\text{H}\{\text{aCLIP-S}(\boldsymbol{c}, \boldsymbol{x}), \max\{\max_{\boldsymbol{r}\in\mathcal{R}} \text{cossim}(\psi(\boldsymbol{c}), \psi(\boldsymbol{r})), 0\}\},$$

where $\text{H}\{\cdot\}$ denotes the harmonic mean of a set. Since our new metrics use data to align CLIP to the downstream task, we categorize them as learning-based [2].

3. Experiments

ReCap We provide details about datasets, baselines, and evaluation metrics in the appendix. In Tab. 1 we show results for MS-COCO and Flickr30k. ReCap outperforms all competitors on our proposed metrics aCLIP-S and RefaCLIP-S on both datasets. On Flickr30k, ReCap attains performance on-par with SmallCap in terms of CIDEr-D and SPICE even though ReCap trains about 1000 times faster with less trainable parameters. On MS-COCO, I-Tuning reaches the highest CIDEr-D and SPICE scores. This gap is due to the fact that I-Tuning trains over 10 times more parameters than ReCap. On both, VizWiz and MSRVTT datasets, ReCap outperforms SmallCap (see Tab. 2). Further, we visualize the joint embedding space of the RN50×64 CLIP encoder without applying our linear alignment for the Flickr30k training set via t-SNE [18] in Fig. 2, left. We find that images and captions are mostly disjoint. However, after applying our linear mapping the two modalities align very well (Fig. 2, right).

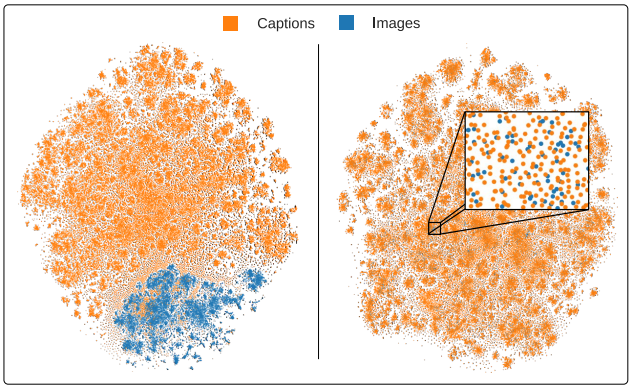

Figure 2. t-SNE visualization of CLIP-embeddings before (left) and after (right) linear alignment on the Flickr30k dataset.

Image Captioning Metrics We elaborate on datasets, baselines, and evaluation metrics in the appendix and report our results in Tab. 3. First, we note that aCLIP-S/RefaCLIP-S consistently outperform CLIP-S/RefCLIP-S [4] from which they were derived. Our aCLIP-S metric achieves the highest correlation with human judgement among all reference-free metrics for both datasets. In the case of reference-based metrics, RefaCLIP-S reaches the highest correlation for Flickr8k-E, while MID reaches the highest correlation for Flickr8k-CF.

4. Conclusion

We advocate for using a linear mapping that can be computed in closed form for two use cases, image captioning and caption evaluation. We introduce ReCap, an efficient retrieval-augmented image-captioning method, which is based on our mapping and requires substantially less training time than other lightweight image-captioning methods. ReCap attains competitive performance to prior lightweight methods on established metrics, effectively reducing training time. We also introduce aCLIP-S and RefaCLIP-S, two new caption evaluation metrics that use our mapping

Table 1. Comparison of different lightweight methods on the MS-COCO test set. We show performance for ReCap and ReCap+DAL. We report mean and standard error for results we computed ourselves. Results for other methods are taken from their respective publications. N/A indicates that a certain metric is not available for a given method.

	MS-COCO				Flickr30k			
Method	CIDEr-D	SPICE	aCLIP-S	RefaCLIP-S	CIDEr-D	SPICE	aCLIP-S	RefaCLIP-S
ClipCap [13]	103.8 ± 1.0	19.9 ± 0.1	46.3 ± 0.2	56.6 ± 0.2	57.0 ± 1.8	15.8 ± 0.3	34.3 ± 0.4	44.0 ± 0.4
I-Tuning$_{\text{Base}}$ [11]	116.7	**21.8**	N/A	N/A	61.5	16.9	N/A	N/A
Prefix-Diffusion [9]	106.3	19.4	N/A	N/A	53.8	14.2	N/A	N/A
SmallCap$_{\text{D=4,Base}}$ [16]	$\mathbf{117.6 \pm 1.0}$	20.0 ± 0.1	46.0 ± 0.2	57.5 ± 0.2	$\mathbf{69.6 \pm 2.1}$	$\mathbf{17.1 \pm 0.3}$	36.8 ± 0.4	46.7 ± 0.4
ReCap (Ours)	108.3 ± 1.0	21.2 ± 0.1	50.4 ± 0.2	60.6 ± 0.2	$\mathbf{68.8 \pm 2.0}$	$\mathbf{17.5 \pm 0.3}$	43.5 ± 0.3	53.4 ± 0.3
ReCap + DAL (Ours)	106.7 ± 1.0	21.2 ± 0.1	$\mathbf{74.7 \pm 0.1}$	$\mathbf{76.2 \pm 0.1}$	$\mathbf{69.5 \pm 2.0}$	$\mathbf{17.3 \pm 0.3}$	$\mathbf{65.7 \pm 0.3}$	$\mathbf{68.2 \pm 0.3}$

Table 2. Comparison of ReCap and SmallCap on the VizWiz and MSRVTT test sets. We report mean and standard error of all methods. CIDEr-D and SPICE on VizWiz are obtained from the official evaluation server. RefaCLIP-S is not available since the VizWiz test set is not public.

	VizWiz			
Method	CIDEr-D	SPICE	aCLIP-S	RefaCLIP-S
ClipCap	48.3	13.4	35.4 ± 0.1	N/A
SmallCap	51.88	13.4	38.4 ± 0.1	N/A
ReCap	**62.3**	**16.7**	$\mathbf{42.7 \pm 0.1}$	N/A
	MSRVTT			
ClipCap	2.0 ± 0.0	10.4 ± 0.0	21.2 ± 0.0	27.5 ± 0.0
SmallCap	31.6 ± 0.2	11.1 ± 0.0	9.2 ± 0.0	7.6 ± 0.3
ReCap	$\mathbf{38.8 \pm 0.2}$	$\mathbf{14.4 \pm 0.0}$	$\mathbf{34.5 \pm 0.0}$	$\mathbf{40.6 \pm 0.0}$

Table 3. Correlation with human judgement measured via Kendall's τ_c for Flickr8k-E and τ_b for Flickr8k-CF both scaled by 100. The variance for the τ estimator only depends on sample size and is 3e-5 for Flickr8k-E and 1e-5 for Flickr8k-CF. † indicates that results were taken from prior work.

Method	Flickr8k-E	Flickr8k-CF
Reference-free		
CLIP-S	51.4	34.3
CLIP+DN	54.0	35.2
aCLIP-S (ours)	**55.1**	**36.2**
Reference-based		
CIDEr-D	43.9	24.6
SPICE	45.0	N/A
RefCLIP-S	53.0	36.4
SoftSPICE†	54.2	N/A
MID†	54.9	**37.3**
CLIP+DN-Ref	55.0	37.0
RefaCLIP-S (ours)	**55.5**	36.7

to adapt CLIP-S and RefCLIP-S, respectively, to a downstream dataset. Our metrics correlate stronger with human judgement than prior CLIP-based metrics and achieve a new state of the art in three out of four categories. In terms of our newly proposed metrics, ReCap outperforms competitors on all tasks. Since the evolution of the field is guided by the metrics that it uses, we hope that this work facilitates research in the direction of image captioning.

References

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas, 2016. Association for Computational Linguistics. 1, 2

[2] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. Learning to evaluate image captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5804–5812. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[3] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, pages 417–434. Springer, 2020. 2

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics, 2021. 1, 2, 3

[5] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.*, 47:853–899, 2013. 2

[6] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 3

[7] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[8] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 2

[9] Guisheng Liu, Yi Li, Zhengcong Fei, Haiyan Fu, Xiangyang Luo, and Yanqing Guo. Prefix-diffusion: A lightweight diffusion model for diverse image captioning. *CoRR*, abs/2309.04965, 2023. 4

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, abs/2304.08485, 2023. 1

[11] Ziyang Luo, Zhipeng Hu, Yadong Xi, Rongsheng Zhang, and Jing Ma. I-tuning: Tuning frozen language models with image for lightweight image captioning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 4

[12] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 1

[13] Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP Prefix for Image Captioning. *CoRR*, abs/2111.09734, 2021. arXiv: 2111.09734. 1, 4

[14] Yassine Ouali, Adrian Bulat, Brais Martínez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15488–15500. IEEE, 2023. 1

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1

[16] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2840–2849. IEEE, 2023. 1, 2, 4

[17] Peter Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 2

[18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 3

[19] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2

[20] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2: 67–78, 2014. 2

[21] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592, 2023. 1

Supplementary Material
Linear Alignment of Vision-language Models for Image Captioning

Acknowledgements

The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects AI-MOTION (LIT-2018-6-YOU-212), DeepFlood (LIT-2019-8-YOU-213), Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), EPILEPSIA (FFG-892171), AIRI FG 9-N (FWF-36284, FWF-36235), AI4GreenHeatingGrids(FFG-899943), INTEGRATE (FFG-892418), ELISE (H2020-ICT-2019-3 ID: 951847), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01). We thank Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), University SAL Labs initiative, FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, GLS (Univ. Waterloo) Software Competence Center Hagenberg GmbH, TÜV Austria, Frauscher Sensonic, Borealis AG, TRUMPF and the NVIDIA Corporation.

Supplementary Material

First, we provide the source code to reproduce all our experiments in Sec. 1. To provide further insights into our method ReCap, we provide additional results on cross-modal retrieval, ablation studies, effect of different data sources, our DAL, and our evaluation as image captioning metric in Sec. 3. Further, we provide more qualitative analysis on retrieved captions after the linear alignment and the effect of synthetic captions in Sec. 4. Sec. 6 gives a rigorous theoretical intuition on the motivation of our linear alignment. Finally, Sec. 5 elaborates on the different hyperparameters we searched, including the retrieval parameter k, the decoding strategy, different vision encoders, generative language models, etc.

1. Source Code

To facilitate reproducibility of our findings, we will make all our code publicly available upon acceptance.

2. Implementation Details

2.1. ReCap

During downstream evaluation of our linear alignment we rely on cosine similarity for retrieval of texts related to an image. Therefore, we evaluate all CLIP vision encoders on cross-modal retrieval tasks in Sec. 3 to find a suitable encoder for ReCap. Based on our findings, we choose RN50×64 [20] as our retrieval model.[1] After embedding images and captions we normalize and center them as suggested by Artetxe et al. [4]. To compute our mapping, we use orthogonal procrustes by default as described by. In certain settings, we use an unconstrained version, i.e., ordinary least squares. We elaborate in Sec. 3 which version we use for the different experiments.

To find the best setting for image captioning, we search over different LMs, decoding strategies, and prompt orderings. We only considered generative LMs that are publicly available on the huggingface hub [58]. Moreover, we search over multiple values of retrieved captions (k). We always search hyperparameters on the validation split of the respective dataset. For more details about hyperparameters, see Sec. 5. We use `faiss` [28] to manage our datastore since it enables efficient storage and retrieval of vectors. Our final setting uses a FLAN-T5-Large [6] with nucleus sampling. To generate captions with FLAN-T5, we explore different prompting strategies and found the strategy proposed in Ramos et al. [48] to work best. Specifically, the used prompt template is "*Similar images show:* $<\text{caption}_1>, \ldots, <\text{caption}_k>$ *This image shows:*".

[1] We take the RN50×64 model from the official repository at https://github.com/openai/CLIP.

Datasets We split the MS-COCO and Flickr30k benchmarks according to Karpathy and Fei-Fei [29] into train, validation, and test splits. For MSRVTT and VizWiz we split according to the official splits [18, 59]. Since VizWiz contains a substantial amount of noise, we filter out all captions for images that suffer from severe quality issues or were rejected by annotators and evaluate the generated test captions on the official evaluation server.[2] For MSRVTT, we employ the same pre-processing pipeline as Ramos et al. [48] and extract four frames from each video and pair them with the ground truth captions. This results in many-to-many correspondences.

Baselines We consider existing methods as lightweight if their trainable parameter count is below 50 M. For MS-COCO and Flickr30k, we compare ReCap to ClipCap [43], I-Tuning [40], SmallCap [48], and Prefix-Diffusion [39]. For MSRVTT and VizWiz, we compare ReCap to SmallCap, since it is the only existing lightweight method that report results on these datasets.

Evaluation Metrics We report metrics commonly used for image captioning, such as CIDEr-D [52] and SPICE [3].[3] We report standard error for all methods we trained ourselves. We do not report error bars for VizWiz since the evaluation server does not provide them. We highlight the best performing methods in boldface throughout the paper and consider two methods to be on-par when their standard errors overlap (68.2% confidence intervals).

2.2. Image Captioning Metrics

Following standard practice of Hessel et al. [21] and Zhou et al. [65], we evaluate our proposed metrics for image captioning by measuring their correlation with human rankings of candidate captions.

Datasets We use the Flickr8k-Expert (Flickr8k-E), and the Flickr8k-Crowdflower [23, Flickr8k-CF] datasets. These datasets provide candidate captions along with human rankings for images of the test set of Flickr8k. We provide additional results for the THumB [30] dataset in Sec. 3.

Baselines We compare our metrics to the current state-of-the-art reference-based and reference-free metrics. In the case of reference-free metrics, we compare to CLIP-score [21], and CLIP+DN [65]. We compare our reference-based metric to RefCLIPScore [21], CLIP+DN-Ref [65], MID [32], and SoftSPICE [37]. For all CLIP+DN variants (reference-based and reference-free) we estimate the mean of both modalities on the respective training dataset, since we usually do not have access to test samples.

Evaluation Metrics To quantify correlation with human judgement, we report Kendall's τ_c for Flickr8k-E and Kendall's τ_b for Flickr8k-CF as done in prior work [21, 65]. The Kendall rank correlation coefficient measures the ordinal association between rankings by humans and the metric. The variance for the τ estimator only depends on sample size and is 3e-5 for Flickr8k-E and 1e-5 for Flickr8k-CF.

3. Additional Results

Cross-modal retrieval We evaluate all publicly available CLIP vision encoders on cross-modal retrieval on the MS-COCO and Flickr30k datasets. We report average recalls and standard error in Tab. 1. We find that larger models improve retrieval performance and, perhaps surprisingly, the RN50×64 encoder outperforms the largest ViT variant in four out of 6 categories when considering image to text retrieval on MS-COCO and Flickr30k. Since ReCap is based on image to text retrieval we select RN50×64 as our retrieval model.

Impact of Linear Alignment We conduct an ablation study where we assess the effect of the linear alignment. To this end, we evaluate a setting where we do not use our linear alignment, which we call ReCap_{ZS}, where ZS stands for zero-shot, since it does not require any training. Further, we distinguish between two types of linear alignment, (i) constrained using orthogonal procrustes (PR), and (ii), unconstrained using ordinary least squares (OLS). Results on the MS-COCO test set are shown in Tab. 2. We observe a substantial performance drop on all metrics for ReCap_{ZS}, showcasing the effectiveness of our linear alignment. The best performing method in terms of CIDEr-D and SPICE is $\text{ReCap}_{\text{OLS}}$, since the unconstrained

[2] https://eval.ai/web/challenges/challenge-page/739/overview

[3] We use the code from https://github.com/tylin/coco-caption.

mapping leads to a stronger alignment with reference captions. The best performance on our learning-based metrics is achieved by ReCap. On one hand we observe the trend that on OLS alignment achieves a better trade-off between rule-based and our learning-based metrics. The PR alignment on the other hand diverges more from reference captions and attains the best performance on our learning-based metrics. Further, the PR alignment leads to higher correlation with human judgement. Thus, we recommend the following criterion for when to deploy which optimization scheme:

- For retrieval-augmented caption generation, use OLS
- For caption evaluation use PR

Training Efficiency We report the training and inference times of different established lightweight image captioning approaches in Tab. 3.

4. Additional Qualitative Analysis

We show some examples for retrieval with and without our linear alignment in Fig. 1. The top row shows the top-k samples for using off-the-shelf CLIP for retrieval, while the bottom row shows retrieval for our aligned CLIP. After the linear alignment, the retrievals fit better to the image. For example, CLIP assigns a high similarity to "*open suitcase*" for the figure in the middle, although the suitcase in the image is closed. Our aligned CLIP does not assign a high similarity to the same caption anymore, and retrieves more appropriate captions.

Figure 1. Sample images and retrieved captions with (bottom) and without (top) our linear alignment to MS-COCO training data. We show three of the closest captions to an image. Images are taken from the MS-COCO validation set.

5. Hyperparameter Search

Effect of different vision encoders We investigate the effect of different vision encoders on the captioning performance of ReCap on the MS-COCO validation set. In this regard, we compare all publicly available encoder variants of CLIP, which comprise ViT-based [8], as well as resnet-based [20] architectures. The best performing model for our retrieval-based image captioning is RN50×64 (see Tab. 4). This corroborates our results for cross-modal retrieval, where RN50×64 outperformed all other encoders Sec. 3.

Top-k retrieval We search over different values for our hyperparameters k on the MS-COCO, Flickr30k, VizWiz, and MSRVTT validation sets. We report results in Tab. 5 and Tab. 6 for MS-COCO, and Flickr30k, respectively. The results for VizWiz and MSRVTT are shown in Tab. 7, and Tab. 8, respectively. For searching over values for k we use greedy decoding, to isolate the effect of the hyperparameter.

Language-model scales We evaluate FLAN-T5 model sizes of 80 M, 250 M, 720 M, 3 B, and 11 B scales. Further, we include decoder-only LMs, such as GPT-2 [46], GPT-J [54], and Llama 7B [51]. The results can be observed in Tab. 9. Our results show that there is not much performance gain going from FLAN-T5-LARGE to FLAN-T5-XXL. We suspect this is due to the design of the prompt which apparently suits FLAN-T5-LARGE particularly well. Surprisingly, even the small variant of FLAN-T5 reaches a CIDEr-D score above 90, which amounts to decent captioning quality.

Our results for decoder-only LMs show that they generally perform worse than encoder-decoder ones. We found that decoder-only models are generally more sensitive to prompt ordering, which was also found in prior works [64]. Perhaps surprisingly, GPT-J outperforms the recently proposed Llama, which reaches performance on-par with GPT-2. Generally, we belive that we could improve performance of larger models by more extensive prompt tuning. However, remarkably, FLAN-T5 performs really well in our setup without the need for extensive prompt tuning.

Different decoding strategies As illustrated by [24], the decoding strategy substantially affects human approval of generated captions. Therefore, we evaluate different decoding strategies, including greedy decoding, sampling, top-k sampling, and nucleus sampling. First, we search over different temperatures τ and number of generated captions l for nucleus sampling [24]. After sampling l captions from the LM, we select the highest scoring one according to our aligned CLIP. To find the best parameters τ and l we set k to the best value we found in the preceeding gridsearch with greedy decoding. Results are reported in Tab. 11, and Tab. 10 for MS-COCO, and Flickr30k, respectively. The results for VizWiz and MSRVTT are shown in Tab. 12, and Tab. 13, respectively.

The results for other decoding schemes are shown in Tab. 14. For greedy decoding we only generate one caption, hence no selection step is required after generation. We use the same temperature as the best nucleus sampling setting for topk and regular sampling. We find that nucleus sampling with $l = 1$ performs close to greedy decoding, however when setting $l = 10$ and using caption selection via our aligned CLIP, we observe a substantial improvement.

Prompt ordering Usually we would provide the captions in the prompt from most-similar to least similar, i.e. the least similar prompt is the most recent in the context. However, one may think the exact opposite ordering might lead to better captioning performance, since the LM might exhibit a form of recency bias. This concerns our setting as well, since the values we found for k are larger than one might expect, e.g., on MS-COCO we found $k = 13$ to perform best. Hence, we provide results for the worst-to-best ordering in Tab. 15. Indeed, we found that different ordering of captions in the prompt leads to different results. Ordering from worst-to-best, i.e. most similar captions appear more recently, leads to an improvement on CIDEr-D score. Therefore, by default, we provide the prompts in the order from worst-to-best in the prompt.

6. Motivation of Linear Alignment

CLIP has been trained to align text with images in a joint embedding space. We want to use the CLIP encoders for retrieval by cosine similarity on an image-captioning task. However, there might be a disparity between the pretraining domain of CLIP and the downstream task. We aim to rectify this by a linear mapping. Our downstream task is retrieval of text embeddings $\boldsymbol{e}_i$ by their corresponding image embeddings $\boldsymbol{f}_i$ using the cosine similarity. Therefore, our objective is

$$\max_{\boldsymbol{W}} \sum_i \text{cossim}(\boldsymbol{e}_i, \boldsymbol{W}\boldsymbol{f}_i). \tag{1}$$

For objective (1) a closed-form solution is unknown. By constraining $\boldsymbol{W}$ to be an orthogonal matrix, however, we obtain equivalence to the least-squares objective because

$$\arg\max_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} \sum_i \text{cossim}(\boldsymbol{e}_i, \boldsymbol{W}\boldsymbol{f}_i) \tag{2}$$

$$= \arg\max_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} \sum_i \frac{\boldsymbol{e}_i^\top \boldsymbol{W}\boldsymbol{f}_i}{\|\boldsymbol{e}_i\|_2 \|\boldsymbol{W}\boldsymbol{f}_i\|_2} \tag{3}$$

$$= \arg\max_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} \sum_i \boldsymbol{e}_i^\top \boldsymbol{W}\boldsymbol{f}_i \tag{4}$$

$$= \arg\min_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} -\sum_i \boldsymbol{e}_i^\top \boldsymbol{W}\boldsymbol{f}_i \tag{5}$$

$$= \arg\min_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} \sum_i (\|\boldsymbol{W}\boldsymbol{f}_i\|_2^2 + \|\boldsymbol{e}_i\|_2^2 - 2\boldsymbol{e}_i^\top \boldsymbol{W}\boldsymbol{f}_i) \tag{6}$$

$$= \arg\min_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} \sum_i (\boldsymbol{f}_i^\top \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{f}_i + \boldsymbol{e}_i^\top \boldsymbol{e}_i - 2\boldsymbol{e}_i^\top \boldsymbol{W}\boldsymbol{f}_i) \tag{7}$$

$$= \arg\min_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} \sum_i (\boldsymbol{W}\boldsymbol{f}_i - \boldsymbol{e}_i)^\top (\boldsymbol{W}\boldsymbol{f}_i - \boldsymbol{e}_i) \tag{8}$$

$$= \arg\min_{\boldsymbol{W}^\top \boldsymbol{W}=\boldsymbol{I}} \sum_i \|\boldsymbol{W}\boldsymbol{f}_i - \boldsymbol{e}_i\|_2^2. \tag{9}$$

Artetxe et al. [4] have pointed out this fact previously. Note that from (3) to (4) and from (5) to (6) the term $\|\boldsymbol{W}\boldsymbol{f}_i\|_2$ can be dropped/added as it appears constant to the optimization objective because $\boldsymbol{W}$ is orthogonal and, therefore, preserves the norm of $\boldsymbol{f}_i$. The solution to this optimization problem is known as orthogonal procrustes [50] and can be written as

$$\boldsymbol{W} = \boldsymbol{V}\boldsymbol{U}^\top, \tag{10}$$

where $\boldsymbol{V}$ and $\boldsymbol{U}$ are the orthogonal matrices of the singular value decomposition of $\boldsymbol{F}^\top \boldsymbol{E} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top, \boldsymbol{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)^\top$.

7. Related Work

Linear Alignment The idea of linearly aligning embedding spaces is a well studied problem in the field of bilinguality [4, 42], geometrical alignment [11, 34, 38], and vision for zero-shot learning [1, 2, 12, 49]. Similar to our approach, Ouali et al. [45] use orthogonal procrustes to align features of CLIP-style models with class labels for few-shot classification. However, their approach is tailored toward the task of classification and does not directly transfer to image captioning. Other works consider image captioning using only text data by training a text decoder for CLIP-style models [17, 35, 44, 56, 62].However, at test-time these approaches still receive images as input, and thus, still suffer from the prevalent mis-alignment. Other approaches adapt the pretraining objective in order to achieve a better alignment in the joint embedding space [13, 15, 26]. However, none of these models are available at the same scale as CLIP.

Retrieval Augmentation The idea of retrieval augmentation has been explored in the realm of language modeling [5, 19, 31], language generation conditioned on images [25, 60, 61], and reinforcement learning [16, 27]. In the realm of image captioning, Ramos et al. [48] leverages retrieval augmentation to reduce the required number of trainable parameters. Ramos et al. [47] extends this idea to multilingual datastores, which enables generation in a certain target language. ReCap also relies on retrieval augmentation, but is much more efficient in terms of training while yielding competitive or even better results.

Lightweight Image Captioning Lightweight captioning aims at reducing the computational complexity for training image captioning models. One line of work is based on knowledge distillation [22] and assumes access to teacher captioning models that are distilled into much smaller scale models [10, 55, 57]. Another line of works leverage parameter-efficient fine-tuning methods to merge visual knowledge into generative LMs via adapter layers [9, 14, 63], cross-attention modules [40, 48], or a mapping network between embedding spaces [41, 43]. Finally, while being lightweight, Kuo and Kira [33] relies on a two-stage training procedure that includes fine-tuning via reinforcement learning [7, 36, 53]. In contrast to ReCap, these methods require end-to-end training.

Table 1. Comparison of different CLIP vision encoders on the cross-modal retrieval task on MS-COCO and Flickr30k. We report average recalls and standard error for all publicly available CLIP vision encoders. Boldface indicates highest average scores.

	MS-COCO					
	IMAGE → TEXT			TEXT → IMAGE		
METHOD	R@1	R@5	R@10	R@1	R@5	R@10
$\text{CLIP}_{\text{RN50}}$	50.2 ± 0.7	74.9 ± 0.6	83.3 ± 0.5	28.4 ± 0.5	52.6 ± 0.5	64.2 ± 0.5
$\text{CLIP}_{\text{RN50x4}}$	52.2 ± 0.7	75.9 ± 0.6	67.5 ± 0.5	31.3 ± 0.5	55.7 ± 0.5	66.5 ± 0.5
$\text{CLIP}_{\text{RN50x16}}$	53.6 ± 0.7	77.9 ± 0.6	85.8 ± 0.5	33.2 ± 0.5	57.0 ± 0.5	67.5 ± 0.5
$\text{CLIP}_{\text{RN50x64}}$	$\mathbf{60.7 \pm 0.7}$	$\mathbf{82.2 \pm 0.5}$	$\mathbf{88.5 \pm 0.5}$	34.3 ± 0.5	59.5 ± 0.5	69.9 ± 0.5
$\text{CLIP}_{\text{ViT-B/32}}$	52.3 ± 0.7	76.0 ± 0.6	84.4 ± 0.5	30.2 ± 0.5	55.1 ± 0.5	66.4 ± 0.5
$\text{CLIP}_{\text{ViT-B/16}}$	52.6 ± 0.7	76.9 ± 0.6	85.0 ± 0.5	32.9 ± 0.5	57.7 ± 0.5	68.1 ± 0.5
$\text{CLIP}_{\text{ViT-L/14}}$	57.0 ± 0.7	80.5 ± 0.6	86.9 ± 0.5	36.1 ± 0.5	60.3 ± 0.5	70.3 ± 0.5
$\text{CLIP}_{\text{ViT-L/14@336px}}$	58.5 ± 0.7	81.3 ± 0.6	88.1 ± 0.5	35.9 ± 0.5	60.4 ± 0.5	70.5 ± 0.5
$\text{ACLIP}_{\text{PR,RN50x64}}$	45.3 ± 0.7	69.7 ± 0.7	79.3 ± 0.6	35.4 ± 0.5	59.4 ± 0.5	70.0 ± 0.5
$\text{ACLIP}_{\text{OLS,RN50x64}}$	33.3 ± 0.7	59.2 ± 0.7	70.2 ± 0.6	$\mathbf{41.5 \pm 0.5}$	$\mathbf{66.9 \pm 0.5}$	$\mathbf{77.0 \pm 0.4}$
$\text{ACLIP}_{\text{IT,RN50x64}}$	33.1 ± 0.7	60.3 ± 0.7	71.3 ± 0.6	31.6 ± 0.5	57.1 ± 0.5	68.4 ± 0.5
	FLICKR30K					
$\text{CLIP}_{\text{RN50}}$	80.8 ± 1.3	95.4 ± 0.7	97.8 ± 0.5	57.9 ± 1.1	83.1 ± 0.8	89.8 ± 0.6
$\text{CLIP}_{\text{RN101}}$	79.2 ± 1.3	94.8 ± 0.7	97.8 ± 0.5	57.5 ± 1.1	81.9 ± 0.8	88.6 ± 0.7
$\text{CLIP}_{\text{RN50x4}}$	83.0 ± 1.2	95.9 ± 0.6	98.2 ± 0.4	61.6 ± 1.1	84.7 ± 0.8	90.1 ± 0.6
$\text{CLIP}_{\text{RN50x16}}$	84.2 ± 1.2	97.0 ± 0.5	99.2 ± 0.3	64.5 ± 1.1	85.9 ± 0.7	91.5 ± 0.6
$\text{CLIP}_{\text{RN50x64}}$	$\mathbf{88.5 \pm 1.0}$	98.3 ± 0.4	99.4 ± 0.2	69.1 ± 1.0	$\mathbf{90.7 \pm 0.6}$	$\mathbf{95.0 \pm 0.4}$
$\text{CLIP}_{\text{ViT-B/32}}$	79.8 ± 1.2	96.3 ± 0.6	98.6 ± 0.4	59.3 ± 1.1	83.7 ± 0.8	90.3 ± 0.6
$\text{CLIP}_{\text{ViT-B/16}}$	83.0 ± 1.2	96.3 ± 0.6	99.3 ± 0.3	63.0 ± 1.1	85.9 ± 0.7	91.8 ± 0.6
$\text{CLIP}_{\text{ViT-L/14}}$	85.7 ± 1.1	98.3 ± 0.4	99.3 ± 0.3	64.8 ± 1.1	87.3 ± 0.7	92.4 ± 0.5
$\text{CLIP}_{\text{ViT-L/14@336px}}$	88.5 ± 1.0	$\mathbf{99.3 \pm 0.3}$	$\mathbf{99.6 \pm 0.2}$	67.0 ± 1.0	88.7 ± 0.7	93.4 ± 0.5
$\text{ACLIP}_{\text{PR,RN50x64}}$	78.5 ± 1.3	95.1 ± 0.7	98.1 ± 0.4	67.0 ± 1.0	89.2 ± 0.6	93.7 ± 0.5
$\text{ACLIP}_{\text{OLS,RN50x64}}$	73.6 ± 1.4	95.0 ± 0.7	97.4 ± 0.5	$\mathbf{70.6 \pm 1.0}$	90.6 ± 0.6	94.0 ± 0.5
$\text{ACLIP}_{\text{IT,RN50x64}}$	67.0 ± 1.5	90.5 ± 0.9	96.4 ± 0.6	62.7 ± 1.0	86.1 ± 0.7	91.8 ± 0.5

Table 2. Ablation study for different methods to compute our linear alignment on the MS-COCO test set. We compare unimodal retrieval (UM), the constrained mapping (PR), unconstrained mapping (OLS), and using no mapping at all (ZS). We report mean and standard error for all settings.

METHOD	CIDER-D	SPICE	ACLIP	REFACLIP-S
RECAP_{UM}	81.9 ± 0.9	16.6 ± 0.1	41.8 ± 0.2	51.5 ± 0.2
RECAP_{ZS}	91.1 ± 0.9	19.1 ± 0.1	47.1 ± 0.2	56.7 ± 0.2
RECAP_{IT}	91.0 ± 0.9	18.7 ± 0.1	46.0 ± 0.2	56.1 ± 0.2
RECAP_{PR}	101.9 ± 1.0	20.4 ± 0.1	$\mathbf{52.5 \pm 0.2}$	$\mathbf{61.6 \pm 0.1}$
$\text{RECAP}_{\text{OLS}}$	$\mathbf{108.3 \pm 1.0}$	$\mathbf{21.2 \pm 0.1}$	50.4 ± 0.2	60.6 ± 0.2

Table 3. Number of parameters, training time, and inference time of ReCap compared to existing lightweight image captioning methods. Inference time is measured in seconds on a subset of 1000 images from the MS-COCO test set on an A100 GPU.

METHOD	$\|\theta\|$	TRAINING	INFERENCE
CLIPCAP	43M	6H (GTX1080)	N/A
PREFIX-DIFFUSION	38.25M	N/A	N/A
I-TUNING	14M	N/A	N/A
SMALLCAP$_{\text{D=4,BASE}}$	1.8M	8H(A100)	$\mathbf{0.19 \pm 0.03}$
RECAP	**1.0M**	$\mathbf{20.3s \pm 1.91}$ **(CPU)**	0.47 ± 0.08

Table 4. Search over all publicly available CLIP vision encoder backbones evaluated on the MS-COCO validation set. We report mean and standard error for all settings. $|\theta|$ denotes the number of trainable parameters.

VISION ENCODER	BLEU@1	BLEU@4	ROUGE-L	CIDER-D	SPICE	$\|\theta\|$
RN50	75.5 ± 0.2	28.0 ± 0.3	56.1 ± 0.2	97.0 ± 0.9	19.7 ± 0.1	1 M
RN101	74.6 ± 0.2	27.7 ± 0.3	56.1 ± 0.2	96.3 ± 0.9	19.4 ± 0.1	262 K
RN50X4	75.4 ± 0.2	28.5 ± 0.3	56.6 ± 0.2	99.2 ± 0.9	19.9 ± 0.1	410 K
RN50X16	76.4 ± 0.2	29.3 ± 0.4	57.0 ± 0.2	102.5 ± 0.9	20.4 ± 0.1	590 K
RN50X64	77.7 ± 0.2	30.5 ± 0.4	58.0 ± 0.2	107.3 ± 1.0	21.2 ± 0.1	1 M
VIT-B/32	75.2 ± 0.2	27.9 ± 0.3	56.0 ± 0.2	96.4 ± 0.9	19.4 ± 0.1	262 K
VIT-B/16	76.2 ± 0.2	29.0 ± 0.3	56.7 ± 0.2	101.2 ± 0.9	20.0 ± 0.1	262 K
VIT-L/14	77.0 ± 0.2	29.9 ± 0.4	57.4 ± 0.2	104.7 ± 1.0	20.6 ± 0.1	590 K
VIT-L/14@336PX	77.4 ± 0.2	30.3 ± 0.4	57.7 ± 0.2	105.8 ± 0.9	20.8 ± 0.1	590 K

Table 5. Hyperparameter Search for k on the MS-COCO validation set for different levels of language abstraction using our semantic mapping computed via OLS. We report mean and standard error for all settings. We select the best k according to CIDEr-D score.

k	BLEU@1	BLEU@4	ROUGE-L	CIDER-D	SPICE
SINGLE CAPTIONS					
10	77.4 ± 0.2	30.4 ± 0.4	57.6 ± 0.2	105.2 ± 1.0	20.9 ± 0.1
11	77.4 ± 0.2	30.4 ± 0.4	57.7 ± 0.2	105.4 ± 1.0	20.9 ± 0.1
12	77.4 ± 0.2	30.3 ± 0.4	57.7 ± 0.2	105.2 ± 1.0	20.9 ± 0.1
13	77.4 ± 0.2	30.5 ± 0.4	57.7 ± 0.2	105.5 ± 1.0	20.8 ± 0.1
14	77.4 ± 0.2	30.5 ± 0.4	57.8 ± 0.2	105.4 ± 1.0	20.8 ± 0.1
15	77.3 ± 0.2	30.5 ± 0.4	57.7 ± 0.2	105.4 ± 1.0	20.9 ± 0.1
16	77.2 ± 0.2	30.4 ± 0.4	57.7 ± 0.2	105.4 ± 1.0	20.8 ± 0.1
17	77.2 ± 0.2	30.2 ± 0.4	57.6 ± 0.2	104.9 ± 1.0	20.9 ± 0.1
ALL CAPTIONS					
1	72.7 ± 0.2	24.8 ± 0.3	53.9 ± 0.2	87.0 ± 0.9	18.0 ± 0.1
2	73.7 ± 0.2	26.4 ± 0.3	54.7 ± 0.2	90.8 ± 0.9	18.2 ± 0.1
3	74.0 ± 0.2	26.4 ± 0.3	54.8 ± 0.2	91.0 ± 0.9	18.2 ± 0.1
4	74.0 ± 0.2	26.6 ± 0.3	55.0 ± 0.2	91.3 ± 0.9	18.5 ± 0.1
5	74.0 ± 0.2	26.9 ± 0.3	55.1 ± 0.2	91.6 ± 0.9	18.4 ± 0.1
LOCALIZED NARRATIVES					
1	55.3 ± 0.3	11.7 ± 0.2	43.1 ± 0.2	45.4 ± 0.6	11.9 ± 0.1
2	54.3 ± 0.3	11.8 ± 0.2	43.0 ± 0.2	48.0 ± 0.7	13.2 ± 0.1
3	53.8 ± 0.3	12.3 ± 0.2	43.0 ± 0.2	50.9 ± 0.7	14.0 ± 0.1
4	53.0 ± 0.3	12.1 ± 0.2	42.7 ± 0.2	51.7 ± 0.7	14.3 ± 0.1
5	52.5 ± 0.3	12.0 ± 0.2	42.6 ± 0.2	52.6 ± 0.7	14.4 ± 0.1
6	52.0 ± 0.3	12.3 ± 0.2	42.6 ± 0.2	53.1 ± 0.7	14.6 ± 0.1

Table 6. Hyperparameter Search for k on the Flickr30k validation set for different levels of language abstraction using our semantic mapping computed via OLS. We report mean and standard error for all settings.

k	BLEU@1	BLEU@4	ROUGE-L	CIDEr-D	SPICE
			SINGLE CAPTIONS		
10	74.8 ± 0.5	26.4 ± 0.7	54.5 ± 0.4	63.6 ± 1.9	15.5 ± 0.3
11	74.7 ± 0.5	26.3 ± 0.7	54.5 ± 0.4	64.4 ± 2.0	15.6 ± 0.3
12	74.4 ± 0.5	26.2 ± 0.7	54.6 ± 0.4	64.6 ± 1.9	15.5 ± 0.3
13	74.2 ± 0.5	26.1 ± 0.7	54.6 ± 0.4	64.4 ± 1.9	15.5 ± 0.3
14	74.6 ± 0.5	26.2 ± 0.7	54.3 ± 0.4	64.4 ± 1.9	15.6 ± 0.3
15	74.3 ± 0.5	26.3 ± 0.7	54.5 ± 0.4	64.8 ± 1.9	15.6 ± 0.3
16	75.0 ± 0.5	26.7 ± 0.7	54.7 ± 0.4	64.6 ± 1.9	15.8 ± 0.3
17	74.5 ± 0.5	26.9 ± 0.7	54.8 ± 0.4	65.5 ± 1.9	15.6 ± 0.3
18	74.9 ± 0.5	26.8 ± 0.7	54.8 ± 0.4	66.2 ± 2.0	15.7 ± 0.3
19	74.4 ± 0.5	26.9 ± 0.7	54.8 ± 0.4	65.6 ± 1.9	15.8 ± 0.3
			ALL CAPTIONS		
1	65.8 ± 0.5	20.3 ± 0.7	49.8 ± 0.4	48.7 ± 1.8	13.4 ± 0.3
2	67.9 ± 0.5	21.5 ± 0.7	50.5 ± 0.5	52.2 ± 1.8	13.9 ± 0.3
3	68.1 ± 0.5	22.0 ± 0.7	51.0 ± 0.4	53.2 ± 1.9	13.7 ± 0.3
4	69.6 ± 0.5	23.0 ± 0.7	51.4 ± 0.4	54.4 ± 1.9	14.1 ± 0.3
5	69.0 ± 0.5	23.0 ± 0.7	51.3 ± 0.4	54.5 ± 1.9	14.2 ± 0.3
			LOCALIZED NARRATIVES		
1	54.2 ± 0.6	9.0 ± 0.4	40.4 ± 0.4	24.4 ± 1.3	8.1 ± 0.2
2	52.6 ± 0.6	8.6 ± 0.4	39.3 ± 0.4	23.3 ± 1.1	8.4 ± 0.2
3	52.5 ± 0.6	9.5 ± 0.4	39.6 ± 0.4	25.4 ± 1.2	8.9 ± 0.2
4	51.7 ± 0.6	9.6 ± 0.4	39.3 ± 0.4	26.0 ± 1.2	9.1 ± 0.2
5	51.9 ± 0.6	9.6 ± 0.4	39.1 ± 0.4	25.6 ± 1.2	9.0 ± 0.2

Table 7. Hyperparameter Search for k on the VizWiz validation set for ReCap with our linear alignment. We report mean and standard error for all settings. We select the best k according to CIDEr-D score.

k	BLEU@1	BLEU@4	ROUGE-L	CIDEr-D	SPICE
1	61.8 ± 0.2	15.5 ± 0.2	43.1 ± 0.2	48.5 ± 0.6	12.1 ± 0.1
2	61.8 ± 0.2	16.5 ± 0.2	44.8 ± 0.2	50.9 ± 0.7	13.1 ± 0.1
3	62.5 ± 0.2	16.9 ± 0.2	45.3 ± 0.2	51.1 ± 0.7	13.0 ± 0.1
4	63.2 ± 0.2	17.5 ± 0.2	45.8 ± 0.2	52.7 ± 0.7	13.0 ± 0.1
5	63.3 ± 0.2	17.5 ± 0.2	45.8 ± 0.2	52.6 ± 0.7	13.1 ± 0.1
6	63.3 ± 0.2	17.6 ± 0.2	45.9 ± 0.2	52.4 ± 0.7	13.0 ± 0.1
7	63.0 ± 0.2	17.5 ± 0.2	45.8 ± 0.2	51.7 ± 0.7	12.9 ± 0.1
8	62.8 ± 0.2	17.5 ± 0.2	45.8 ± 0.2	51.6 ± 0.7	12.8 ± 0.1
9	62.9 ± 0.2	17.5 ± 0.2	45.9 ± 0.2	51.3 ± 0.7	12.9 ± 0.1
10	62.1 ± 0.2	17.0 ± 0.2	45.5 ± 0.2	50.3 ± 0.6	12.8 ± 0.1

Table 8. Hyperparameter Search for k on the MSRVTT validation set for ReCap with our linear alignment. We report mean and standard error for all settings. We select the best k according to CIDEr-D score.

k	BLEU@1	BLEU@4	ROUGE-L	CIDEr-D	SPICE
3	26.9 ± 0.1	4.8 ± 0.0	25.7 ± 0.1	36.6 ± 0.4	14.2 ± 0.1
4	26.9 ± 0.1	4.8 ± 0.0	25.7 ± 0.1	36.6 ± 0.4	14.2 ± 0.1
5	27.1 ± 0.1	4.9 ± 0.0	25.8 ± 0.1	36.7 ± 0.4	14.1 ± 0.1
6	27.1 ± 0.1	4.9 ± 0.0	25.8 ± 0.1	36.4 ± 0.4	14.0 ± 0.1
7	27.0 ± 0.1	4.9 ± 0.0	25.9 ± 0.1	36.4 ± 0.3	13.9 ± 0.1
8	27.0 ± 0.1	4.9 ± 0.0	25.9 ± 0.1	36.7 ± 0.4	13.8 ± 0.1

Table 9. Comparison of different language models on the MS-COCO validation set. We report mean and standard error for all settings.

Model	BLEU@1	BLEU@4	ROUGE-L	CIDEr-D	SPICE
Encoder-Decoder					
FLAN-T5-SMALL	63.9 ± 0.3	23.3 ± 0.3	55.0 ± 0.2	93.9 ± 1.0	20.5 ± 0.1
FLAN-T5-BASE	72.5 ± 0.2	27.1 ± 0.3	56.7 ± 0.2	100.0 ± 0.9	20.7 ± 0.1
FLAN-T5-LARGE	77.7 ± 0.2	30.5 ± 0.4	58.0 ± 0.2	107.3 ± 1.0	21.2 ± 0.1
FLAN-T5-XL	76.1 ± 0.2	29.4 ± 0.4	56.7 ± 0.2	104.7 ± 0.9	20.8 ± 0.1
FLAN-T5-XXL	77.1 ± 0.2	30.2 ± 0.4	57.4 ± 0.2	107.0 ± 1.0	21.0 ± 0.1
Decoder-only					
GPT-2	64.9 ± 0.3	24.1 ± 0.3	49.5 ± 0.2	86.8 ± 0.9	19.1 ± 0.1
GPT-J 6B	71.1 ± 0.3	29.1 ± 0.4	51.4 ± 0.2	97.5 ± 1.0	19.6 ± 0.1
Llama 7B	61.5 ± 0.3	23.1 ± 0.3	49.3 ± 0.2	86.4 ± 0.9	19.5 ± 0.1

Table 10. Comparison of different values for temperature of nucleus sampling on the Flickr30k validation set for $k = 18$

Temperature	Samples	BLEU@1	BLEU@4	ROUGE-L	CIDEr-D	SPICE
1.0	1	74.8 ± 0.5	26.8 ± 0.7	54.6 ± 0.4	65.0 ± 1.9	15.8 ± 0.3
0.1	10	75.2 ± 0.5	27.5 ± 0.7	55.2 ± 0.4	68.7 ± 2.0	16.5 ± 0.3
0.3	10	74.5 ± 0.5	26.6 ± 0.7	55.2 ± 0.4	68.4 ± 1.9	16.8 ± 0.3
0.5	10	73.8 ± 0.5	25.6 ± 0.7	54.6 ± 0.4	68.4 ± 2.1	17.0 ± 0.3
0.1	20	75.3 ± 0.5	27.1 ± 0.7	55.2 ± 0.4	68.7 ± 1.9	16.5 ± 0.3
0.3	20	74.4 ± 0.5	26.6 ± 0.7	55.2 ± 0.4	69.3 ± 2.0	16.9 ± 0.3
0.5	20	73.4 ± 0.5	25.2 ± 0.7	54.6 ± 0.4	68.3 ± 2.0	17.3 ± 0.3
0.1	30	75.5 ± 0.5	27.5 ± 0.7	55.3 ± 0.4	68.7 ± 2.0	16.6 ± 0.3
0.3	30	74.2 ± 0.5	26.4 ± 0.7	55.4 ± 0.4	68.9 ± 2.0	17.2 ± 0.3
0.5	30	72.9 ± 0.5	24.4 ± 0.7	54.4 ± 0.4	67.7 ± 2.0	17.3 ± 0.3

Table 11. Comparison of different values for temperature of nucleus sampling on the MS-COCO validation set for $k = 13$.

TEMPERATURE	SAMPLES	BLEU@1	BLEU@4	ROUGE-L	CIDER-D	SPICE
0.0	N/A	77.4 ± 0.2	30.5 ± 0.4	57.7 ± 0.2	105.5 ± 1.0	20.8 ± 0.1
0.1	10	77.7 ± 0.2	30.5 ± 0.4	58.0 ± 0.2	107.3 ± 1.0	21.2 ± 0.1
0.3	10	77.3 ± 0.2	29.9 ± 0.4	57.9 ± 0.2	106.8 ± 0.9	21.4 ± 0.1
0.5	10	76.5 ± 0.2	29.0 ± 0.3	57.3 ± 0.2	104.5 ± 0.9	21.3 ± 0.1
0.1	20	77.6 ± 0.2	30.4 ± 0.4	57.9 ± 0.2	107.2 ± 1.0	21.2 ± 0.1
0.3	20	77.2 ± 0.2	29.7 ± 0.3	57.8 ± 0.2	106.2 ± 0.9	21.4 ± 0.1
0.5	20	76.4 ± 0.2	28.6 ± 0.3	57.1 ± 0.2	103.9 ± 0.9	21.4 ± 0.1
0.1	30	77.6 ± 0.2	30.4 ± 0.4	57.9 ± 0.2	107.1 ± 0.9	21.2 ± 0.1
0.3	30	77.1 ± 0.2	29.5 ± 0.3	57.7 ± 0.2	106.1 ± 0.9	21.4 ± 0.1
0.5	30	76.4 ± 0.2	28.3 ± 0.3	57.1 ± 0.2	103.3 ± 0.9	21.6 ± 0.1

Table 12. Comparison of different values for temperature of nucleus sampling on the VizWiz validation set for $k = 4$.

TEMPERATURE	SAMPLES	BLEU@1	BLEU@4	ROUGE-L	CIDER-D	SPICE
0.0	N/A	63.2 ± 0.2	17.5 ± 0.2	45.8 ± 0.2	52.7 ± 0.7	13.0 ± 0.1
0.1	10	64.5 ± 0.2	17.9 ± 0.2	46.3 ± 0.2	54.7 ± 0.7	13.6 ± 0.1
0.3	10	64.9 ± 0.2	18.2 ± 0.2	46.5 ± 0.2	56.3 ± 0.7	14.1 ± 0.1
0.5	10	64.9 ± 0.2	18.1 ± 0.2	46.5 ± 0.2	56.7 ± 0.7	14.3 ± 0.1
0.1	20	64.5 ± 0.2	18.0 ± 0.2	46.3 ± 0.2	54.8 ± 0.7	13.6 ± 0.1
0.3	20	65.1 ± 0.2	18.3 ± 0.2	46.7 ± 0.2	56.6 ± 0.7	14.3 ± 0.1
0.5	20	65.1 ± 0.2	18.2 ± 0.2	46.5 ± 0.2	57.1 ± 0.7	14.6 ± 0.1
0.1	30	64.6 ± 0.2	18.0 ± 0.2	46.3 ± 0.2	55.0 ± 0.7	13.7 ± 0.1
0.3	30	65.2 ± 0.2	18.3 ± 0.2	46.7 ± 0.2	56.9 ± 0.7	14.3 ± 0.1
0.5	30	64.9 ± 0.2	18.1 ± 0.2	46.7 ± 0.2	58.0 ± 0.7	14.7 ± 0.1

Table 13. Comparison of different values for temperature of nucleus sampling on the MSRVTT validation set for $k = 5$.

TEMPERATURE	SAMPLES	BLEU@1	BLEU@4	ROUGE-L	CIDER-D	SPICE
0.0	N/A	27.1 ± 0.1	4.9 ± 0.0	25.8 ± 0.1	36.7 ± 0.4	14.1 ± 0.1
0.1	10	24.8 ± 0.1	4.4 ± 0.0	25.8 ± 0.1	37.4 ± 0.4	14.7 ± 0.1
0.3	10	24.9 ± 0.1	4.2 ± 0.0	25.6 ± 0.1	38.2 ± 0.4	14.8 ± 0.1
0.5	10	24.7 ± 0.1	4.1 ± 0.0	25.3 ± 0.1	37.9 ± 0.4	14.6 ± 0.1
0.1	20	24.7 ± 0.1	4.3 ± 0.0	25.7 ± 0.1	37.3 ± 0.4	14.7 ± 0.1
0.3	20	24.8 ± 0.1	4.2 ± 0.0	25.6 ± 0.1	38.0 ± 0.4	14.7 ± 0.1
0.5	20	24.6 ± 0.1	4.0 ± 0.0	25.3 ± 0.1	38.3 ± 0.4	14.6 ± 0.1
0.1	30	24.7 ± 0.1	4.3 ± 0.0	25.8 ± 0.1	37.3 ± 0.4	14.7 ± 0.1
0.3	30	24.7 ± 0.1	4.2 ± 0.0	25.6 ± 0.1	38.1 ± 0.4	14.7 ± 0.1
0.5	30	24.5 ± 0.1	4.0 ± 0.0	25.3 ± 0.1	38.1 ± 0.4	14.6 ± 0.1

Table 14. Search over different decoding paradigms for captioning on the MS-COCO validation set. We report mean and standard error for all settings. Sampling-based decoding strategies use a temperature of $\tau = 0.1$.

DECODING	BLEU@1	BLEU@4	ROUGE-L	CIDER-D	SPICE
SAMPLING	67.9 ± 0.2	21.0 ± 0.3	51.6 ± 0.2	80.7 ± 0.8	19.3 ± 0.1
TOPK	67.9 ± 0.2	20.8 ± 0.3	51.5 ± 0.2	80.9 ± 0.8	19.4 ± 0.1
GREEDY	77.4 ± 0.2	30.5 ± 0.4	57.7 ± 0.2	105.5 ± 1.0	20.8 ± 0.1
NUCLEUS, $l = 1$	77.4 ± 0.2	30.4 ± 0.4	57.8 ± 0.2	105.5 ± 1.0	20.8 ± 0.1
NUCLEUS	77.7 ± 0.2	30.5 ± 0.4	58.0 ± 0.2	107.3 ± 1.0	21.2 ± 0.1

Table 15. Comparison of different orderings for exemplars in the prompt on the MS-COCO validation set. We report mean and standard error for all settings.

ORDERING	BLEU@1	BLEU@4	ROUGE-L	CIDER-D	SPICE
WORST-TO-BEST	77.7 ± 0.2	30.5 ± 0.4	58.0 ± 0.2	107.3 ± 1.0	21.2 ± 0.1
BEST-TO-WORST	77.4 ± 0.2	30.4 ± 0.4	57.7 ± 0.2	105.9 ± 1.0	21.0 ± 0.1

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 819–826. IEEE Computer Society, 2013. 5

[2] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 2927–2936. IEEE Computer Society, 2015. 5

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 382–398. Springer, 2016. 2

[4] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas, 2016. Association for Computational Linguistics. 1, 5

[5] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 2206–2240. PMLR, 2022. 5

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. 1

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10575–10584. Computer Vision Foundation / IEEE, 2020. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[9] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. MAGMA - multimodal augmentation of generative models through adapter-based finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2416–2428. Association for Computational Linguistics, 2022. 5

[10] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1408–1418. IEEE, 2021. 5

[11] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 5

[12] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2121–2129, 2013. 5

[13] Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet Thuong Tran, Fei Tang, Hubert Ramsauer, D P Kreil, Michael K Kopp, Günter Klambauer, Angela Bitto-Nemling, and Sepp Hochreiter. CLOOB: Modern hopfield networks with infoLOOB outperform CLIP. In *Advances in Neural Information Processing Systems*, 2022. 5

[14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010, 2023. 5

[15] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 5

[16] Anirudh Goyal, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adria Puigdomenech Badia, Arthur Guez, Mehdi Mirza, Peter C Humphreys, Ksenia Konyushova, et al. Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning*, pages 7740–7765. PMLR, 2022. 5

[17] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! learning visual tasks using only language data. *CoRR*, abs/2211.09778, 2022. 5

[18] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, pages 417–434. Springer, 2020. 2

[19] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 3929–3938. PMLR, 2020. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1, 3

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics, 2021. 2

[22] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 5

[23] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.*, 47:853–899, 2013. 2

[24] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 4

[25] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 23369–23379. IEEE, 2023. 5

[26] Christina Humer, Vidya Prasad, Marc Streit, and Hendrik Strobelt. Understanding and comparing multi-modal models: Exploring the latent space of clip-like models (clip, cyclip, cloob) using inter-modal pairs. *6th Workshop on Visualization for AI Explainability*, 2023. 5

[27] Peter C. Humphreys, Arthur Guez, Olivier Tieleman, Laurent Sifre, Theophane Weber, and Timothy P. Lillicrap. Large-scale retrieval for reinforcement learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 5

[28] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3): 535–547, 2019. 1

[29] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, 2017. 2

[30] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3464–3478. Association for Computational Linguistics, 2022. 2

[31] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 5

[32] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual information divergence: A unified metric for multimodal generative models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[33] Chia-Wen Kuo and Zsolt Kira. HAAV: hierarchical aggregation of augmented views for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11039–11049. IEEE, 2023. 5

[34] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. In *10th IEEE International Conference on Computer Vision (ICCV 2005), 17-20 October 2005, Beijing, China*, pages 1482–1489. IEEE Computer Society, 2005. 5

[35] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In *The Eleventh International Conference on Learning Representations*, 2023. 5

[36] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, pages 121–137. Springer, 2020. 5

[37] Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6377–6390. Association for Computational Linguistics, 2023. 2

[38] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T. Freeman. SIFT flow: Dense correspondence across different scenes. In *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III*, pages 28–42. Springer, 2008. 5

[39] Guisheng Liu, Yi Li, Zhengcong Fei, Haiyan Fu, Xiangyang Luo, and Yanqing Guo. Prefix-diffusion: A lightweight diffusion model for diverse image captioning. *CoRR*, abs/2309.04965, 2023. 2

[40] Ziyang Luo, Zhipeng Hu, Yadong Xi, Rongsheng Zhang, and Jing Ma. I-tuning: Tuning frozen language models with image for lightweight image captioning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2, 5

[41] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 5

[42] Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3992–4006. Association for Computational Linguistics, 2022. 5

[43] Ron Mokady, Amir Hertz, and Amit H. Bermano. ClipCap: CLIP Prefix for Image Captioning. *CoRR*, abs/2111.09734, 2021. arXiv: 2111.09734. 2, 5

[44] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4055–4063. Association for Computational Linguistics, 2022. 5

[45] Yassine Ouali, Adrian Bulat, Brais Martínez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15488–15500. IEEE, 2023. 5

[46] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. 4

[47] Rita Ramos, Bruno Martins, and Desmond Elliott. Lmcap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1635–1651. Association for Computational Linguistics, 2023. 5

[48] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2840–2849. IEEE, 2023. 1, 2, 5

[49] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2152–2161. JMLR.org, 2015. 5

[50] Peter Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 5

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. 4

[52] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society, 2015. 2

[53] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society, 2015. 5

[54] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, 2021. 4

[55] Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Minivlm: A smaller and faster vision-language model. *CoRR*, abs/2012.06946, 2020. 5

[56] Junyang Wang, Ming Yan, Yi Zhang, and Jitao Sang. From association to generation: Text-only captioning by unsupervised cross-modal mapping. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 4326–4334. ijcai.org, 2023. 5

[57] Ning Wang, Jiangrong Xie, Hang Luo, Qinglin Cheng, Jihao Wu, Mingbo Jia, and Linlin Li. Efficient image captioning for edge devices. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 2608–2616. AAAI Press, 2023. 5

[58] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 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, 2020. Association for Computational Linguistics. 1

[59] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2

[60] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Ming-Yu Liu, Yuke Zhu, Mohammad Shoeybi, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *CoRR*, abs/2302.04858, 2023. 5

[61] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 39755–39769. PMLR, 2023. 5

[62] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, Jae Sung Park, Ximing Lu, Prithviraj Ammanabrolu, Rowan Zellers, Ronan Le Bras, Gunhee Kim, and Yejin Choi. Multimodal knowledge alignment with reinforcement learning. *CoRR*, abs/2205.12630, 2022. 5

[63] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *CoRR*, abs/2303.16199, 2023. 5

[64] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 12697–12706. PMLR, 2021. 4

[65] Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser-Nam Lim. Test-time distribution normalization for contrastively learned visual-language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2