SELF-SUPERVISED OPEN-ENDED CLASSIFICATION WITH SMALL VISUAL LANGUAGE MODELS

Mohammad Mahdi Derakhshani[1], **Ivona Najdenkoska**[1]
Cees G. M. Snoek[2], **Marcel Worring**[+], **Yuki M. Asano**[2]
University of Amsterdam
Amsterdam, the Netherlands

ABSTRACT

We present Self-Context Adaptation (SeCAt), a self-supervised approach that unlocks few-shot abilities for open-ended classification with small visual language models. Our approach imitates image captions in a self-supervised way based on clustering a large pool of images followed by assigning semantically-unrelated names to clusters. By doing so, we construct a training signal consisting of interleaved sequences of image and pseudo-caption pairs and a query image, which we denote as the 'self-context' sequence. Based on this signal the model is trained to produce the right pseudo-caption. We demonstrate the performance and flexibility of SeCAt on several multimodal few-shot datasets, spanning various granularities. By using models with approximately 1B parameters we outperform the few-shot abilities of much larger models, such as Frozen and FROMAGe. SeCAt opens new possibilities for research and applications in open-ended few-shot learning that otherwise requires access to large or proprietary models.

1 INTRODUCTION

Language models have demonstrated fascinating emergent abilities, particularly in-context learning (Brown et al., 2020; Wei et al., 2022a). This represents the ability to solve few-shot learning tasks without any gradient-based updates. Recently, such models have evolved from natural language processing domain to visual language models (VLMs) (Tsimpoukelli et al., 2021; Alayrac et al., 2022). Yet, such models heavily rely on incorporating very large, proprietary language backbones, ranging from 7 up to 70 billion parameters, making them impractical for specific downstream tasks. Interestingly, in-context learning abilities have not been yet observed in small-scale models, even for solving open-ended image classification tasks. One reason is that these models rely profoundly on semantic priors created during the pre-training (Wei et al., 2023b). Larger models, by contrast, override these priors, allowing them to learn directly from input-label mappings presented in context. Since the pre-training strategies of both small and large-scale VLMs are similar, we hypothesize that the mechanisms for in-context learning should be present in small models as well.

We start with a pre-trained image captioning model, intending to "teach" it to capture input-label mappings in context. To avoid any manual curation of such mappings, we define our approach in a self-supervised manner. We perform clustering of an unlabelled image dataset, followed by assigning semantically-unrelated names as cluster labels. The usage of such names for clusters gives flexibility to our method because any word can be used for learning the mappings in a prompt. Then, we construct pseudo-captions by using a template "This is a + *cluster name*", which have either random or nonsensical meanings w.r.t the image content. After getting such captions, we construct the so-called *self-context* in a self-supervised manner, which contains interleaved image-caption pairs as context and a query image. We adapt the pre-trained model with mini-batches of these self-contexts, where the model is optimized to generate the caption for the query image given the context sequence. This defines our lightweight procedure, which we name *Self-Context Adaptation (SeCAt)*, and is illustrated in Figure 1.

[1]Shared first authorship. The authors can change the order for their purposes. [2]Shared last-authorship; order random. Corresponding authors: {m.m.derakhshani, i.najdenkoska}@uva.nl

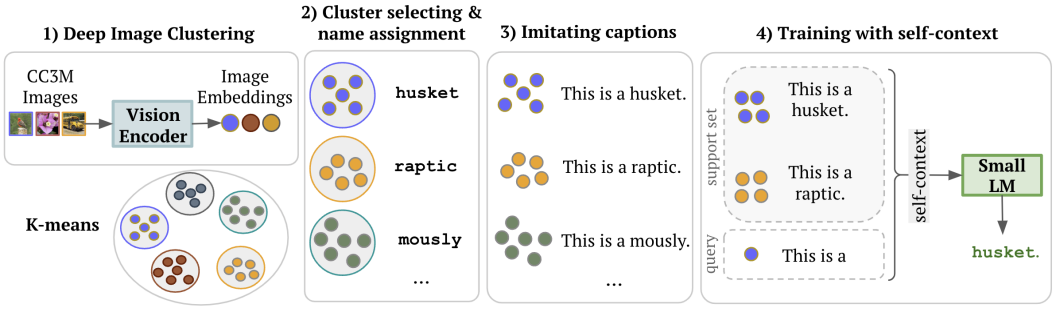

Figure 1: The SeCAt method consists of four steps: First, the image embeddings are extracted with a vision encoder, followed by deep image clustering. Next is the selection of clusters and assigning arbitrary names to them. Then, the assigned names are used to imitate image captions for each image in the selected cluster. The last step is the self-context adaptation of the small language model, by using the previously generated image-caption pairs.

At inference time, we keep the vision and language backbones entirely frozen and we prompt the model with interleaved contexts to perform open-ended image classification. For this, we employ the multimodal few-shot datasets proposed in (Tsimpoukelli et al., 2021). Furthermore, to test the ability of the model to deal with different levels of task granularity, we also evaluate our approach on semantically-easy and hard few-shot tasks based on five common vision datasets. With this, we show that the flexibility of constructing self-contexts allows us to control the difficulty and granularity of the few-shot tasks. Last but not least, we show that SeCAt can turn even small visual language models of the order of 1 billion parameters into strong in-context learners, without any supervised fine-tuning.

To summarize, we contribute in three major aspects: *Conceptual:* We present an efficient framework for unlocking in-context learning for open-ended few-shot learning. *Methodological:* We define a self-supervised adaptation procedure to learn an in-context template for small VLMs. *Empirical:* We conduct extensive experiments on several multimodal few-shot datasets, ranging from coarse to fine-grained tasks, and show that we achieve better performance compared to the larger visual-language counterparts.

2 Methodology

To enable open-ended few-shot learning via in-context mechanisms, we propose a self-supervised adaptation technique that mimics the final in-context learning objective but does not rely on any labeled or captioned data.

At a high level, our method clusters a large pool of images to identify highly coherent groups and assigns them names that are meant to not necessarily fit or describe the content. This noisy set of images and names is then used for adapting the model in a manner that simulates in-context learning. Our method allows for controlling the context difficulty by sampling items from distant or close clusters and by doing so it allows the final model to work well even for fine-grained few-shot learning. In the next sections, we will first formally state the problem, then describe the procedure for generating the self-supervised image-caption pairs, then we will outline the construction of self-context training samples and how we vary their difficulty, as well as the final training procedure.

Problem statement. Open-ended few-shot learning aims to generate the correct caption t_q corresponding to a query image x_q, given pairs of images x_s and captions t_s in a support-set $s \in \mathcal{S}$, handled by a VLM denoted as f:

$$f(\{(x_s, t_s)\}_{s \in \mathcal{S}}, x_q) = t_q. \tag{1}$$

Following the standard few-shot learning paradigm, for the model to "learn" from the context, the support-set $\mathcal{S}$ contains a similar image as the query. In the case of utilizing an LM as a decoder, the task is "open-ended", *i.e.*, t_q must be obtained via text generation, and not via classification into a fixed set of labels. Naturally, we can train a VLM with this objective, be it that this requires access to a set of paired image-text data, as we can see from Eq. 1. Instead of obtaining supervised sets

of image-caption pairs, we propose to mimic this data using self-supervision and use the generated image-text pairs to finetune the VLM.

Model overview. The architecture that we use is based on image captioning encoder-decoder models. Note that in these models, such as ClipCap (Mokady et al., 2021), f is the model that first embeds an image with a vision encoder Ψ and then maps it into the representation space of a language model (LM), *i.e.*, f=LM($\Psi(x)$). To perform this mapping, it uses a mapping function implemented as a simple multi-layer perceptron, which outputs the visual embeddings as a visual prefix for the language model.

2.1 Generating Image-Pseudo Caption Pairs

Imitating image labels. Let $h : x \rightarrow c$ define the human annotation process of classifying an image x in a dataset $\mathcal{X}$ into class $c \in \mathcal{C}$ of a classification system $\mathcal{C}$. We replace h by a composition of two unsupervised functions, $h \approx c \odot m$. The first component c, first clusters the dataset $\mathcal{X}$ in a self-supervised manner. For this, we utilize the visual embeddings obtained by a visual encoder Ψ and cluster the whole dataset, defined as:

$$c(x) = K\text{-means}[\{\Psi(x')\}_{x' \in \mathcal{X}}](x), \tag{2}$$

where K is the number of clusters and the resulting output of c indicates the cluster ID for a given image. Next, we assign each ID to an arbitrary label to obtain the paired data.

Imitating image captions. To arrive at pairings of captions to a given image cluster c, we utilize a vocabulary of words $w \in \mathcal{V}$ (we show that a list of random names suffices for this). Next, we utilize the VLM f for the cluster name assignment, i.e. the matching step. To match the words with clusters, we embed with Ψ one exemplar image per cluster, namely the cluster centroid, and embed the vocabulary words into their language model token-space using the tokenizer-embedding function τ. Note that now, both $\Psi(x)$ and $\tau(w)$ are in the same embedding space, so we can simply construct a similarity matrix $\mathbf{S} \in \mathbb{R}^{K \times |\mathcal{V}|}$ by computing their cosine-similarities:

$$\mathbf{S} = \text{sim}(\Psi(x), \tau(w)). \tag{3}$$

Finally, we match each image cluster with a word embedding by using the Kuhn-Munkres (Hungarian) algorithm (Kuhn, 1955) to minimize the overall cost. The algorithm takes this output and yields the assigned word given a cluster ID. Afterward, the captions are imitated by converting these cluster names into "This is a + *cluster name*" captions (note that other templates are also possible) and are paired with all images belonging to the particular cluster.

2.2 Self-Context Construction

To construct an interleaved sequence of self-context samples, we randomly pick images according to their cluster membership, during the mini-batch construction. By choosing the level of similarity between two or more clusters, from which the support set is constructed, we can control the difficulty of the classification problem. For a given cluster k, we then sample items (x_i, t_i) s.t. $c(x_i)$=k, which represent an image-caption pair belonging to the self-context, illustrated in the Figure A.1.

We also vary the difficulty of the few-shot tasks depending on the proximity between cluster centroids. This means that if two clusters are far away from each other, they create an *easy* self-context. If they are close they create a *hard* self-context since the image samples from closer clusters have potentially more visual similarities between each other, rather than distant clusters.

2.3 Mixed Self-Context Learning & Inference

Given the image-caption mappings (x_s, t_s) as a self-context, and the query image x_q, the learning process is performed by optimizing the cross-entropy loss, while generating the query caption t_q:

$$\mathcal{L} = H(f(\{(x_s, t_s)\}_{s \in \mathcal{S}'}, x_q)|t_q). \tag{4}$$

Note that the loss function uses the constructed self-context as a single data point. To encourage generalization to different context lengths with one model we perform *mixed* self-context learning,

		Real-Name miniImageNet				Open-Ended miniImageNet			
		2-way		5-way		2-way		5-way	
Methods	#params	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ClipCap (Mokady et al., 2021)	1.3B	0.0	0.0	0.02	0.0	0.0	0.0	0.0	0.01
Frozen (Tsimpoukelli et al., 2021)	7B	33.7	66.0	14.5	33.8	53.4	58.9	51.1	**58.5**
FROMAGe (Koh et al., 2023)	6.7B	31.0	50.4	17.5	30.7	27.8	49.8	16.3	19.5
SeCAt (Ours)	1.3B	**85.7**	**83.2**	**68.6**	**58.0**	**87.4**	**85.6**	**68.0**	41.9
OpenFlamingo (Awadalla et al., 2023)	9B	62.0	95.9	45.3	91.2	45.2	63.4	15.0	56.9

Table 1: Baselines comparison on 2- and 5-way Real-Name miniImageNet and Open-Ended mini-ImageNet in accuracy(%). OpenFlamingo (Awadalla et al., 2023) is considered an upper-bound.

		Easy split				Hard split			
		2-way		5-way		2-way		5-way	
Methods	#params	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
ClipCap (Mokady et al., 2021)	1.3B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FROMAGe (Koh et al., 2023)	6.7B	30.0	50.1	13.8	28.3	28.6	46.6	10.0	23.5
SeCAt (Ours)	1.3B	**81.3**	**65.2**	**70.5**	**49.7**	**63.8**	**52.6**	**34.7**	**26.2**
OpenFlamingo (Awadalla et al., 2023)	9B	53.3	98.9	37.8	98.8	39.9	90.3	25.9	78.0

Table 2: Generalization from easy-to-hard on the 2- and 5-way Easy vs Hard dataset splits in accuracy(%). Note that OpenFlamingo (Awadalla et al., 2023) is considered an upper-bound.

where we randomly vary the context length within a batch. This means that we change the number of samples in the context by taking into account 2-way and j-shot tasks alternately, where $j \in \{1, 3, 5\}$.

At inference time, we keep the full model entirely frozen, and we test its ability to digest new in-context sequences. We consider previously unseen few-shot tasks, which also have a support set as a context, and a query sample to evaluate the performance. The model completes the sentence for each query sample in an autoregressive manner. To obtain the final output, we use beam-search to sample from the language model given the sequence of context samples.

3 Experiments

3.1 Results & Discussion

Baseline comparison. In multimodal few-shot learning scenarios, the model needs to learn the connection between visual concepts and words by observing only a few demonstrations. The experiments in Table 1, measure to what extent our SeCAt approach can perform such binding with LMs of 1.3B parameters. Our approach outperforms models that are up to 5× larger, such as Frozen (Tsimpoukelli et al., 2021) and FROMAGe (Koh et al., 2023). This shows that small models can indeed be adapted to be good few-shot learners for few-shot learning in a fast and efficient manner. We view OpenFlamingo (Awadalla et al., 2023) as an upper-bound of our approach since it employs 5× more parameters and is pre-trained on web-scraped interleaved sequences of images and text, which directly helps in-context learning abilities. While OpenFlamingo employs 5× more parameters and trains on extensive datasets such as LAION2B (Schuhmann et al., 2022) (with 2B image-text pairs) and Multi-modal C4 (Zhu et al., 2023) (with 104M combined image-text samples), our method bypasses such extensive pre-training by leveraging our unique self-supervised approach.

Generalization from easy-to-hard. The flexibility of our approach, to select clusters with a particular distance and label them in a self-supervised manner, allows us to handle different levels of granularity of few-shot tasks. In Table 2, we demonstrate the performance on easy and hard splits, which are illustrated in Figure 2 and Section A.2. As expected, it is easier for the model to adjust to the easy-split settings, compared to the hard-split. Similarly as in Table 1, our approach outperforms FROMAGe (Koh et al., 2023), across all settings, even though it is using a notably smaller LM.

Qualitative analysis. In Figure 3, we show an example of a 2-way 1-shot task, with an interleaved image-caption sequence and a query image. It can be seen that SeCAt successfully binds visual

Figure 2: Examples of the restructured datasets to obtain the easy and hard few-shot tasks. The top row illustrates a 5-way 1-shot task from the easy split, with a shot per dataset. The bottom row depicts a 5-way 1-shot task from the hard split, where all shots are selected from one dataset.

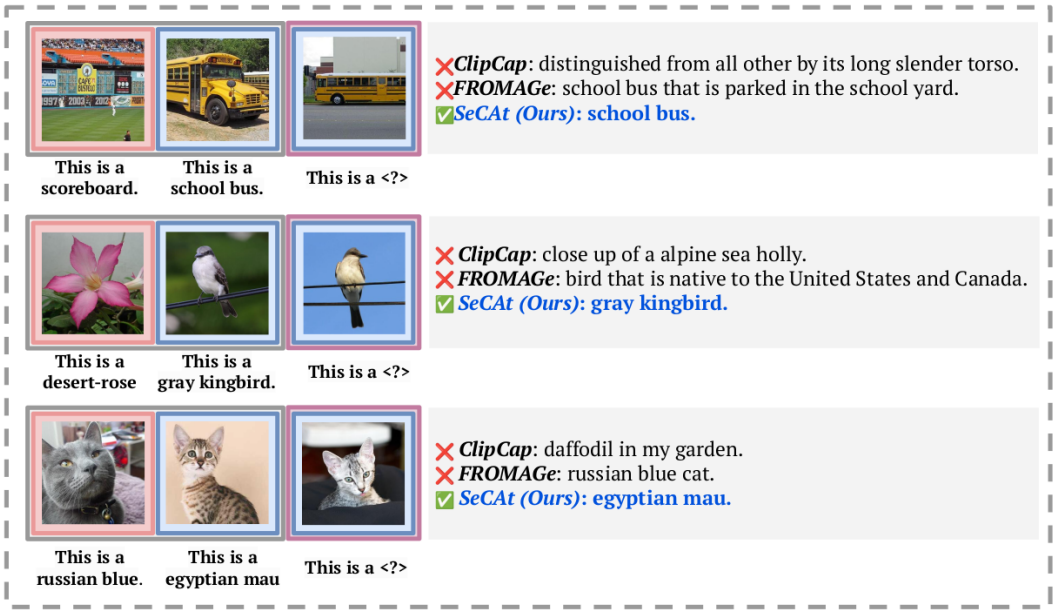

Figure 3: SeCAT model at inference time: Comparison to two other baselines, ClipCap and FROMAGe, on a 2-way 1-shot tasks from Real-Names miniImageNet dataset.

concepts in the image to the relevant words, and can produce the expected output. However, ClipCap generates an incorrect caption, not related to the query image, showing the lack of in-context learning ability in small VLMs without SeCAt. Interestingly, FROMAGe can capture the concept of *school bus* or *bird* as predictions, but it is also excessively verbose. This essentially means that it is leveraging its semantic priors from the image captioning pre-training and not entirely adapting to the context sequence. We provide additional qualitative comparisons in the appendix.

4 Conclusion

We present Self-Context Adaptation (SeCAt), a self-supervised learning method that enhances small visual language models for open-ended few-shot learning by clustering unlabelled images and assigning them unrelated names to mimic image captions. It generates sequences of self-contexts for the language model, enabling it to recognize patterns and dependencies within the context. Our experiments confirm that SeCAt is efficient in data and training resources, making advanced multimodal few-shot learning more accessible.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016. 11

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances on Neural Information Processing Systems*, 2022. 1, 10

Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Advances on Neural Information Processing Systems*, 2020a. 10

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *International Conference on Learning Representations*, 2020b. 10

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. OpenFlamingo, March 2023. URL https://doi.org/10.5281/zenodo.7733589. 4

Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, 2017. 10

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 11

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances on Neural Information Processing Systems*, 2020. 1, 10

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018. 10

Stephanie CY Chan, Adam Santoro, Andrew K Lampinen, Jane X Wang, Aaditya Singh, Pierre H Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent few-shot learning in transformers. In *Advances on Neural Information Processing Systems*, 2022. 10

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 10

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics*, 2019. 10

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 11

Kirill Gavrilyuk, Mihir Jain, Ilia Karmanov, and Cees G M Snoek. Motion-augmented self-training for video recognition at smaller scale. In *ICCV*, 2021. 10

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 10

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 10

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 10

Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 10

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 10

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 11

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2016. 11

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023. 4, 10, 13

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3, 12

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Annual Meeting of the Association for Computational Linguistics*, 2021. 10

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 10

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Conference on Computer Vision and Pattern Recognition*, 2022. 10

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Annual Meeting of the Association for Computational Linguistics*, 2022. 10

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3, 4, 10

Ivona Najdenkoska, Xiantong Zhen, and Marcel Worring. Meta learning to bridge vision and language models for multimodal few-shot learning. In *International Conference on Learning Representations*, 2023. 10

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 11

Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Conference on Computer Vision and Pattern Recognition*, 2018. 10

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Conference on Computer Vision and Pattern Recognition*, 2012. 11

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 *International Conference on Machine Learning*, 2021. 10, 11

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 10

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021. 10

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances on Neural Information Processing Systems*, 2022. 10

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances on Neural Information Processing Systems*, 2022. 4

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018. 10

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022. 10

Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P Xing, and Zhiting Hu. Progressive generation of long text with pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 10

Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q Tran, David R So, Siamak Shakeri, Xavier Garcia, Huaixiu Steven Zheng, Jinfeng Rao, Aakanksha Chowdhery, et al. Transcending scaling laws with 0.1% extra compute. *arXiv preprint arXiv:2210.11399*, 2022. 10

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances on Neural Information Processing Systems*, 2021. 1, 2, 4, 10, 12

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*, 2020. 10

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 11

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 10

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021. 10

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a. 1

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022b. 10

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*, 2023a. 10

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023b. 1

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. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. 11

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition*, 2010. 11

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, 2016. 10

Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *Conference on Computer Vision and Pattern Recognition*, 2016. 10

Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In *Conference on Empirical Methods in Natural Language Processing*, 2022. 10

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023. 4

A Appendix

A.1 Related work

Few-shot Learning in Language Models. Large language models have garnered substantial attention within the NLP community (Brown et al., 2020; Chan et al., 2022; Chowdhery et al., 2022; Dai et al., 2019; Tan et al., 2020; Yang et al., 2022; Wei et al., 2023a) due to their capacity to generate extensive text as well as their remarkable in-context capabilities. Achieving this, often requires scaling transformer-based models (Rae et al., 2021; Smith et al., 2022; Chowdhery et al., 2022), augmenting pre-training data (Hoffmann et al., 2022), and advanced loss functions (Wei et al., 2021; Tay et al., 2022). The in-context learning paradigm was first introduced by GPT3 (Brown et al., 2020) as a *training-free* learning framework for few-shot learning. Numerous works have further explored this ability and showcased that it makes it easier to incorporate outside knowledge into language models by changing the context and templates (Liu et al., 2021; Wei et al., 2022b; Lu et al., 2021). Yet, the emergent in-context learning ability comes with the cost of a huge number of parameters and a large-scale pre-training dataset. For instance, GPT3 consists of 175B parameters and is trained on approximately 45TB of text data. Another recent work, introduced symbolic tuning Wei et al. (2023a) by also using semantically-unrelated words. However, they only focus on language-based tasks. Different from these works, we propose an algorithm that unlocks in-context learning in small visual language models for open-ended few-shot learning.

Multimodal Few-shot Learning. Recent advancements in vision and language have arisen with the emergence of large language models (Radford et al., 2021; Ramesh et al., 2021; Saharia et al., 2022; Alayrac et al., 2022; Jia et al., 2021; Hao et al., 2022; Najdenkoska et al., 2023; Wang et al., 2022). We highlight Flamingo (Alayrac et al., 2022), FROMAGe (Koh et al., 2023), and ClipCap (Mokady et al., 2021) as notable examples. In these works, the in-context ability emerges by scaling up the number of transformer parameters, which has previously proven effective in various NLP tasks. Additionally, several methods, including Flamingo, FROMAGe, MetaLM (Hao et al., 2022), and KOSMOS-1 (Huang et al., 2023), incorporate interleaved sequences of images and captions during training. This approach simulates few-shot learning scenarios, enabling large language models to capture patterns among multiple image-caption pairs within a single sequence, thereby facilitating few-shot learning. It is important to note that ClipCap (Mokady et al., 2021) does not exhibit the in-context learning mechanism as it is not trained on interleaved images and captions. Similar to FROMAGe and Flamingo, our method benefits from interleaved sequences of images and text during training, while we differ in language model size, pre-training dataset size, and the use of distinct loss functions during the adaptation phase. Despite our focus on small-scale visual language models, we still enable in-context learning capabilities for multimodal few-shot learning problems.

Unsupervised Pseudo-label Generation. Generating pseudo-labels by clustering has proven effective in unsupervised representation learning (Asano et al., 2020b; Caron et al., 2018; Ji et al., 2019; Van Gansbeke et al., 2020; Xie et al., 2016; Yang et al., 2016). This approach involves using pseudo labels in the visual domain for tasks such as image representation learning (Caron et al., 2018; Bojanowski & Joulin, 2017; Noroozi et al., 2018), image segmentation (Melas-Kyriazi et al., 2022), and video understanding (Asano et al., 2020a; Gavrilyuk et al., 2021). A self-labeling method is proposed in Asano et al. (2020b), driven by k-means and repurposed to learn a shared set of labels between audio and text modalities. Inspired by this work, we propose a self-supervised approach using k-means clustering to assign semantically-unrelated words as labels to the visual clusters and then imitate interleaved sequences of image-caption pairs based on these labels.

A.2 Experimental setup

Datasets. To pre-train an image captioning model and to perform the clustering part, we use the *Conceptual Captions (CC3M)* dataset (Sharma et al., 2018), which consists of 3M pairs of images and captions, web-scrapped and post-processed. At the inference stage, we employ multimodal few-shot datasets (Tsimpoukelli et al., 2021), namely *Real-Names miniImageNet* and *Open-Ended miniImageNet*, each one with 1, 3 and 5 shots, with 2 and 5-way tasks. The evaluation setting is similar to MetaICL (Min et al., 2022) which also investigates in-context abilities but only for text classification.

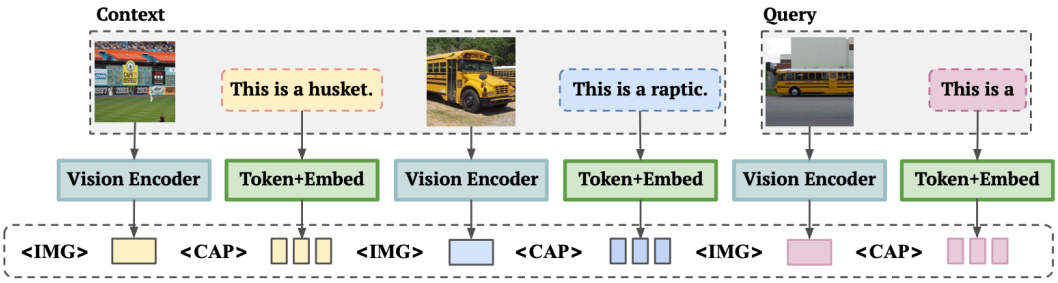

Figure A.1: The self-context is represented as a sequence of interleaved pairs of images and pseudo-captions. It uses special tokens such as <IMG> and <CAP> to denote the position of the elements in the sequence and is used as input to the language model to complete the sentence for the query image. Note that this is an example of a 2-way 1-shot self-context sequence.

Additionally, to test the ability of our approach to generalize across fine-grained and coarse-grained settings, we create semantically *easy* and *hard* datasets. In particular, we reorganize existing datasets, namely *OxfordPets* (Parkhi et al., 2012), *Flowers102* (Nilsback & Zisserman, 2008), *Food101* (Bossard et al., 2014), *CUBS-200* (Wah et al., 2011) and *SUN397* (Xiao et al., 2010). For the semantically-easy split, given an n-ways k-shots scenario, we randomly choose n datasets from the pool of these five datasets. Subsequently, from each selected dataset, we randomly select a single class to constitute the n-ways setting. For the semantically-hard split, we randomly select one dataset, followed by the selection of n classes from that chosen dataset. Finally, from the chosen classes, we randomly select k-image samples. We provide more details about the construction of the easy and hard-splits in the appendix. The splits will be released to foster further study.

Implementation details. The language backbone of our model is based on the GPT-family of models, namely GPT-Neo model (Gao et al., 2020), and the smaller versions, GPT2-small and GPT2-medium. We utilize the vision encoder from the pre-trained CLIP ViT-B/32 model (Radford et al., 2021) for our model's visual backbone. To ensure that the model correctly pays attention to the image and caption during training, we add special tokens <IMG> and <CAP> in the prompt before the image and caption respectively (see Figure A.1). We have found this to be particularly useful for in-context learning because it helps the language model to focus on attending to the correct image and text within the interleaved prompt sequence. To implement the deep clustering stage, we use the Faiss library (Johnson et al., 2019), particularly the k-means algorithm with 10 iterations. The full implementation is in PyTorch and HuggingFace (Wolf et al., 2020) and will be publicly released. We provide additional details on the implementation and hyperparameters in the appendix.

Training details. Our models are trained using mixed-precision with Bfloat16 (Abadi et al., 2016). In the image captioning pre-training stage, we use a batch size of 160 over 370,000 iterations and 3 A6000 GPUs. Furthermore, we use the AdamW optimizer (Kingma & Ba, 2016) with a learning rate of $2e$-5 and a warmup of 5000 steps. We set the visual prefix length to 5 and the word embedding dimension to 2048. During the self-context adaptation stage, we only fine-tune the language backbone with a small learning rate of $5e$-6 for 50 epochs and keep all other components fixed.

Evaluation criteria. We evaluate our approach in an open-ended fashion, by measuring the accuracy(%) of generating the words which match the ground-truth.

A.3 Ablations

In the next sections, we ablate our method on the Real-Name miniImageNet dataset using 2-way and 5-ways in both 1-shot and 5-shot settings.

Effect of self-context difficulty. Our method is sufficiently flexible to vary the difficulty of the self-context construction. We can use cluster centroids, in close proximity or further apart from each other, to influence the semantics of the chosen visual concepts within the self-context. We consider three different settings by computing L2 distances between all centroids. The *hard* setting takes the most similar 5%, the *easy* setting takes the least similar 5%, and the *varying* setting shuffles the clusters from both the hard and easy settings. As can be observed from Table A.1a, the hard setting performs considerably worse than the other two, as the model deals with images clustered closely

(a) Effect of self-context difficulty.

	2-way		5-way	
difficulty	1-shot	5-shot	1-shot	5-shot
hard	32.6	39.4	14.9	8.6
easy	82.2	81.8	52.5	29.8
varying	**85.7**	**83.2**	**68.6**	**58.0**

(b) Influence of semantically-unrelated names.

	2-way		5-way	
vocabulary	1-shot	5-shot	1-shot	5-shot
nonsense	77.2	69.7	55.7	10.3
numbers	81.6	54.8	49.4	24.9
nouns	**85.7**	**83.2**	**68.6**	**58.0**

(c) Matching names to cluster centroids.

	2-way		5-way	
matching	1-shot	5-shot	1-shot	5-shot
random	81.8	83.2	**68.7**	40.7
cost-based	**85.7**	**83.2**	68.6	**58.0**

(d) Benefit of mixed self-context training.

	2-way		5-way	
setting	1-shot	5-shot	1-shot	5-shot
single-task	73.3	25.1	35.2	3.6
mixed-task	**85.7**	**83.2**	**68.6**	**58.0**

(e) Impact of language model size.

	2-way		5-way	
LM	1-shot	5-shot	1-shot	5-shot
$GPT2_{small}$	26.9	54.3	37.5	33.1
$GPT2_{medium}$	56.2	64.2	42.4	41.7
GPT-Neo	**85.7**	**83.2**	**68.6**	**58.0**

(f) Generalization on different prompt templates.

	2-way		5-way	
Template	1-shot	5-shot	1-shot	5-shot
"A photo of a"	73.0	70.3	45.0	43.2
"On this picture there is a"	72.5	67.8	58.2	39.2
"This is a"	**85.7**	**83.2**	**85.7**	**58.0**

Table A.1: Ablations. We ablate the key components of our method, namely (a) Effect of self-context difficulty, (b) Influence of semantically-unrelated names, (c) Matching names to cluster centroids, (d) Benefit of mixed self-context training, (e) Impact of language model size, and (f) Generalization on different prompt templates. Evaluations are done on the 2- and 5-way Real-Name miniImageNet with the best model from Table 1.

together with limited variability. For both the easy and varying settings the performance increases. We conclude that our approach benefits from varying the proximity between cluster centroids.

Influence of semantically-unrelated names. For the selection of the semantically-unrelated names used for labeling the clusters and then generating the pseudo-captions of images, we consider either nonsense words, random numbers, or random nouns. The nonsense words are taken using a nonsense-word generator[1], similar to Tsimpoukelli et al. (2021). The random numbers and nouns are generated in a similar manner and are semantically-*unrelated* to the clustered images. Table A.1b shows the performance per vocabulary choice, across different few-shot settings. The random nouns yield better performance than the random numbers and nonsense names. Even though the cluster names are unrelated to the images in the cluster, the model still achieves satisfactory performance. This suggests that any word embedding is good enough for the model to learn since it views them as mere symbols helpful for learning a self-context pattern.

Matching names to cluster centroids. The impact of the name-matching techniques is explored in Table A.1c, where we compare random cluster-name matching and cost-based matching. In the random cluster-name matching variant, the name embeddings are randomly assigned to cluster centroids. The cost-based matching variant utilizes the Kuhn-Munkres (Hungarian) algorithm Kuhn (1955), which aims to find the minimal distance between cluster centroids and name embeddings. The cost-based matching approach yields better performance, which means that SeCAt benefits from a more informed manner of cluster naming.

Benefit of mixed self-context training. To evaluate the influence of varying self-context length, we consider two adaptation strategies. The first strategy, denoted as single-task, is simply using a fixed number of samples in the self-context across all mini-batches, where we consider only 2-way 1-shot tasks. The second strategy is the mixed self-context training, where we randomly vary the number

[1] https://www.soybomb.com/tricks/words/

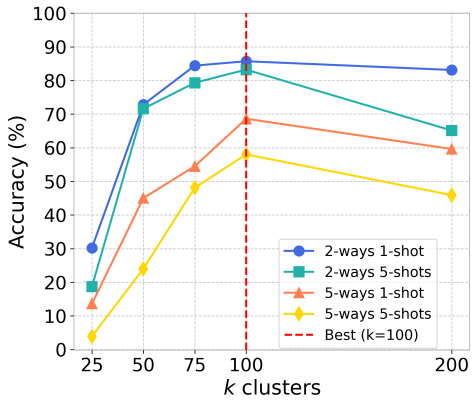

Figure A.2: Influence of varying numbers of clusters for generating the pairs of images and pseudo-captions. The accuracy increases up to 100 clusters.

of samples by using 2-way and j-shot tasks, where $j \in \{1, 3, 5\}$. Comparing the two strategies in Table A.1d reveals that mixed self-context training consistently outperforms the single one by a considerable margin, especially when the number of shots increases. This is mainly attributed to the fact that the mixed training paradigm lets the model observe different lengths of the self-context sequences.

Impact of language model size. To investigate the impact of the language model, we replace the GPT-Neo backbone (1.3B parameters) with its smaller alternatives GPT2-medium (355M parameters) and GPT2-small (124M parameters) and report results in Table A.1e. Naturally, the best performance is obtained with the largest variant, but the two smaller alternatives also show satisfying results, especially if we take into account the considerable difference in size.

We looked at the training times required for each variant of the language backbone. The best version of our approach using GPT-Neo can be trained in just 14 hours, unlike larger variants which require more than a day of training (e.g. FROMAGe (Koh et al., 2023)). Moreover, training the smaller variants is even faster: 6 hours for GPT2-small and 11 hours for GPT2-medium. This time efficiency is crucial when rapid model adaptation is necessary or when access to large models and computational resources is limited.

Generalization on different prompt templates. We adapt the model using the common prompt "This is a + *label*" and report results based on it. To demonstrate the robustness of our model to other prompts at inference time, we introduce alternative prompt templates. Particularly, we use: "A photo of a + *label*" and "On this picture, there is a + *label*". The performance, as presented in Table A.1f, affirms our method's strong generalization across varied prompt templates, negating the possibility of overfitting to a specific prompt.

Influence of the varying number of clusters. The number of clusters can be tuned depending on the fine-graininess of the problem at hand. We evaluate our best setting by using different numbers of k clusters, where $k \in \{25, 50, 75, 100, 200\}$. As can be seen in Figure A.2, we observe a consistent increase in the performance of up to 100 clusters. We assume that, as the miniImageNet evaluation datasets are not fine-grained enough, the performance slightly starts degrading for $k = 200$.

Limitations. Our work aims to unlock in-context learning in small visual language models for open-ended few-shot learning. It achieves the necessary capacities to some degree, but it can benefit from extending the evaluation on more complicated tasks which can give a clearer picture of possible applications. However, it is already able to achieve good performance on open-ended few-shot learning, which can be easily extended to other open-ended vision-language tasks, such as image captioning and visual question answering as future work.

A.4 Additional Qualitative Evaluation

In the next section, we provide additional qualitative comparisons between SeCAt-trained model and two other baselines, namely ClipCap and FROMAGe. We show a few successful cases in Figure A.3 and also failure cases in Figure A.4.

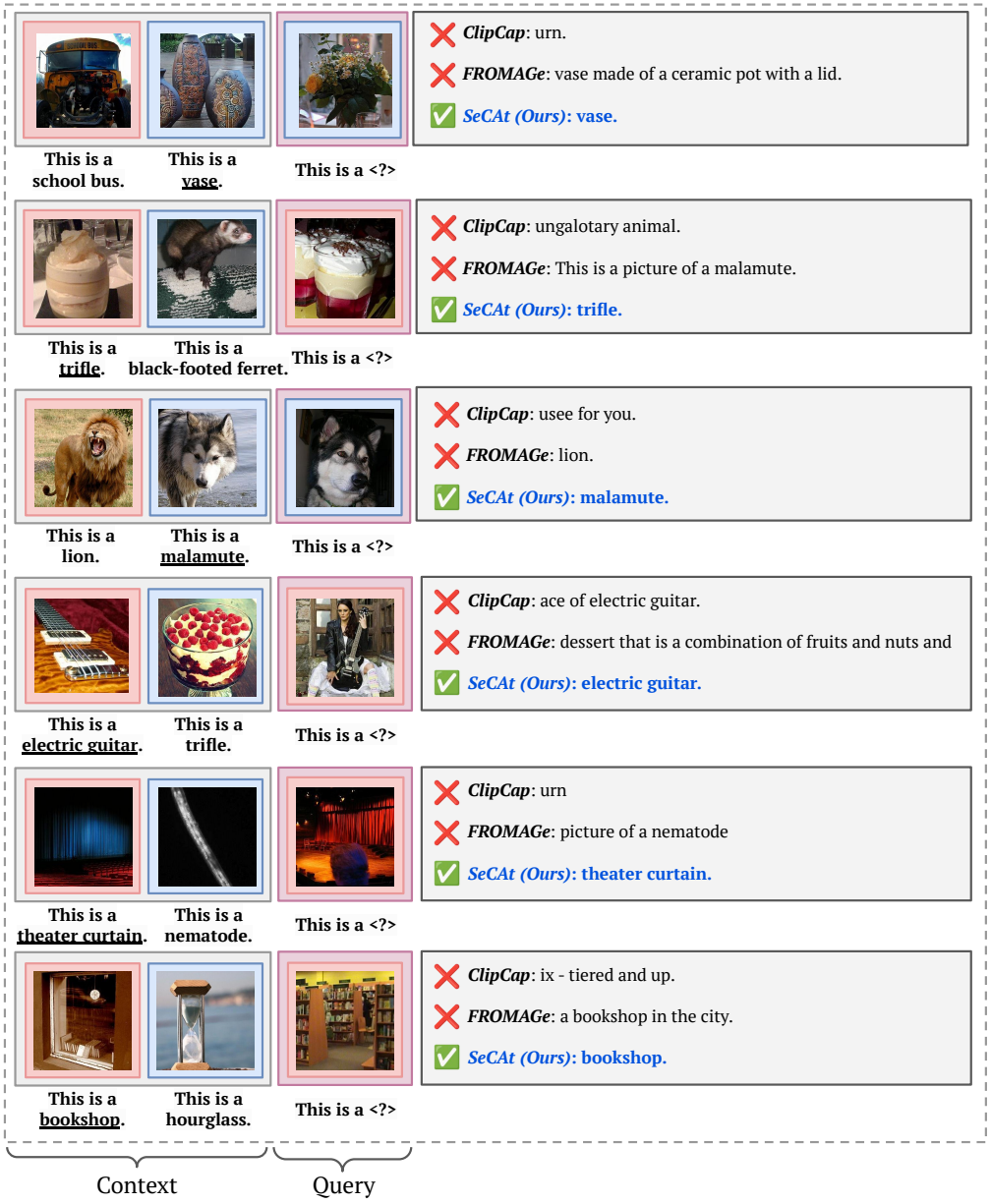

Figure A.3: Qualitative comparison between SeCAt-trained model and two other baselines, namely ClipCap and FROMAGe, on a 2-way 1-shot task from Real-Names miniImageNet, showing successful cases of SeCAt.

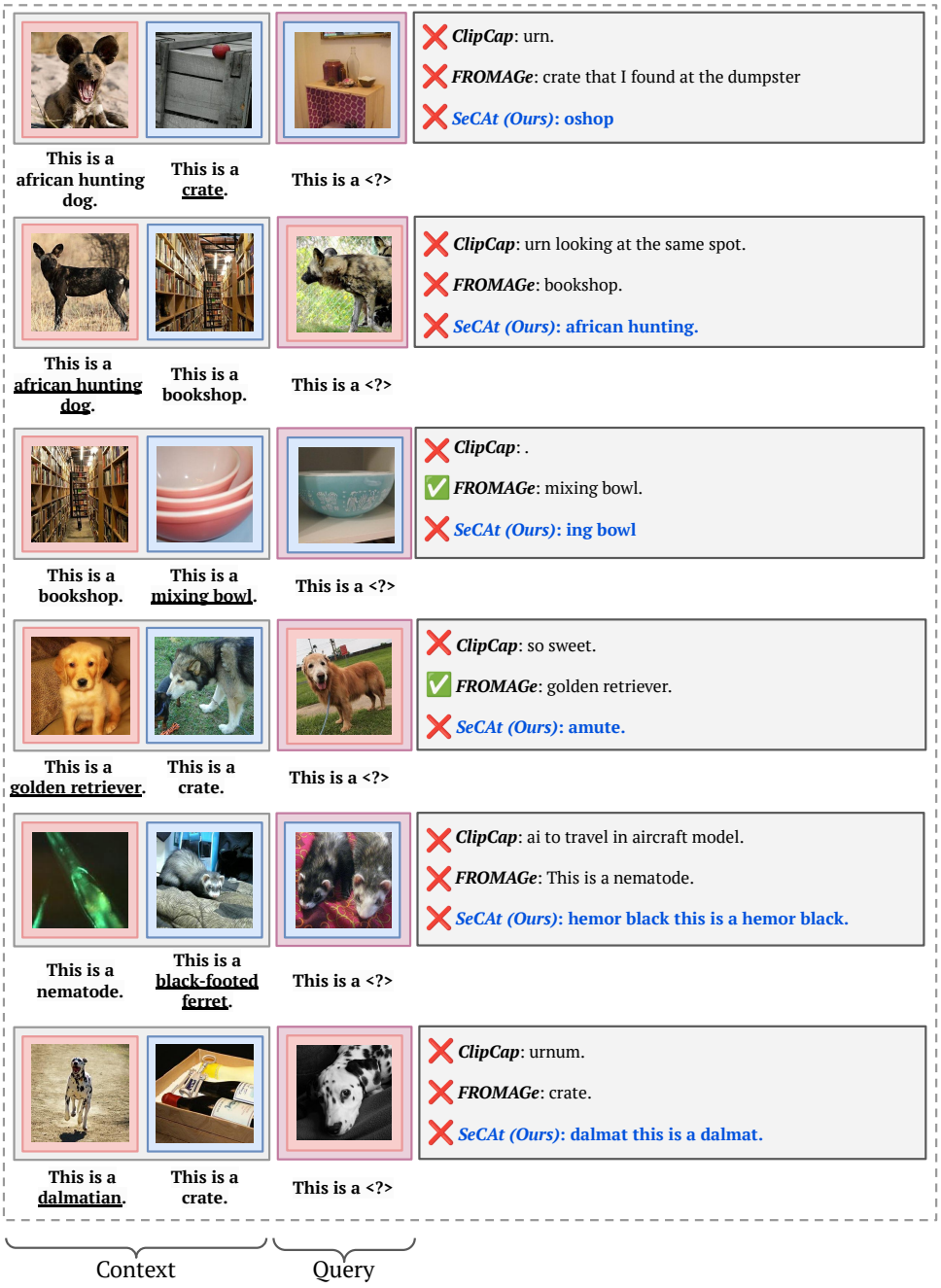

Figure A.4: Qualitative comparison between SeCAt-trained model and two other baselines, namely ClipCap and FROMAGe, on a 2-way 1-shot task from Real-Names miniImageNet, showing failure cases of SeCAt.