Enhancing Conceptual Understanding in Multimodal Contrastive Learning through Hard Negative Samples

Philipp J. Rösch[1], **Norbert Oswald**[1], **Michaela Geierhos**[1], and **Jindřich Libovický**[2]
[1]Bundeswehr University Munich, Germany
[2]Faculty of Mathematics and Physics, Charles University, Czech Republic
{philipp.roesch,norbert.oswald,michaela.geierhos}@unibw.de
libovicky@ufal.mff.cuni.cz

Abstract

Current vision-language models leveraging contrastive learning often face limitations in developing fine-grained conceptual understanding. This is due to random negative samples during pretraining, causing almost exclusively very dissimilar concepts to be compared in the loss function. Consequently, the models struggle with fine-grained semantic differences. To address this problem, we introduce a novel pretraining method incorporating synthetic hard negative text examples. The hard negatives replace terms corresponding to visual concepts, leading to a more fine-grained visual and textual concept alignment. Further, we introduce InpaintCOCO, a new challenging dataset for assessing the fine-grained alignment of colors, objects, and sizes in vision-language models. We created the dataset using generative inpainting from COCO images by changing the visual concepts so that the images no longer match their original captions. Our results show significant improvements in fine-grained concept understanding across various vision-language datasets, including our InpaintCOCO dataset.

1 Introduction

Recent advancements in vision-language (VL) modeling have demonstrated the effectiveness of contrastive learning in various multimodal tasks (Radford et al., 2021; Jia et al., 2021; Yao et al., 2021). However, this training method does not provide sufficient training signals for several important visual concepts (Zhao et al., 2023). We attributed it to the objective function's use of random and, therefore, too dissimilar negative samples, which prevents the model from learning fine-grained semantic representations of the concepts.

Therefore, we propose a novel approach to address the issue of poorly represented concepts in contrastive learning. We introduce a mechanism to incorporate hard negative samples into the contrastive learning loss. Specifically, we generate synthetic hard negative samples by substituting keywords in the captions of original image-text pairs, disrupting the alignment between the image content and its description.

This paper presents three key contributions:

(i) We present a novel method for using hard negative samples in the contrastive learning objective, allowing the model to focus on refining its understanding of concepts.

(ii) By introducing hard negative samples in the language component, we compel the model to learn proper visual and language alignment. Our approach improves multimodal performance, although it operates exclusively on the language side of the model.

(iii) To evaluate the model from the visual perspective, we create a challenge set with over 1,260 adversarial examples by using generative image inpainting. This dataset serves as a comprehensive benchmark, allowing us to assess the model's ability to validate its conceptual understanding. This is because the image was created in a standardized setting in which only a small part was changed.

In this work, we conduct extensive evaluations of four basic concepts – color, object, location, and size. These concepts were selected as examples to demonstrate the effectiveness and robustness of our proposed approach in capturing nuanced semantic relations, but it is important to note that the choice of concepts is flexible and can be tailored to specific applications. Furthermore, our methodology is easy to construct, requiring only minimal domain expertise and the simple usage of regular expressions. This study shows that simple tweaks in contrastive learning can significantly enhance multimodal understanding and model performance.

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 102–115
August 16, 2024 ©2024 Association for Computational Linguistics

I_1

T_1^{pos}	T_2^{neg}	T_3^{neg}	T_1^{hn}
a white cat sits under a black open umbrella.	a person standing on a loading platform next to a train.	a man leans near the door of a creatively painted tour bus.	a **brown** cat sits under a black open umbrella.

Figure 1: Classical contrastive learning approaches use (I_1, T_1^{pos}) as positive pairs in combination with negative samples like T_2^{neg} and T_3^{neg} to learn an image-text alignment. A bag of words (e.g., nouns) is often sufficient to extract the correct text that matches a given image, resulting in only broad concepts learned. We also use hard negatives like T_1^{hn} so that fine-grained semantic concepts are learned for visual and textual alignment.

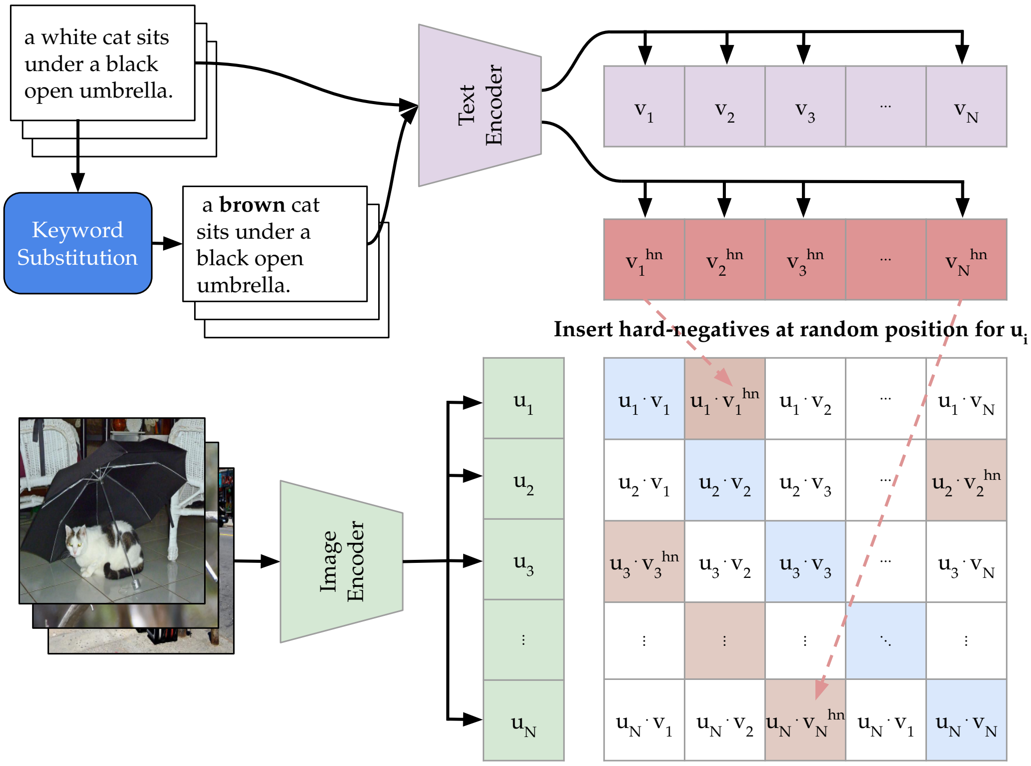

a white cat sits under a black open umbrella.
Keyword Substitution
a **brown** cat sits under a black open umbrella.
Text Encoder
v1 v2 v3 ··· vN
v1hn v2hn v3hn ··· vNhn
Insert hard-negatives at random position for ui
Image Encoder
u1 u2 u3 ⋮ uN
u1·v1 u1·v1hn u1·v2 ··· u1·vN
u2·v1 u2·v2 u2·v3 ··· u2·v2hn
u3·v3hn u3·v2 u3·v3 ··· u3·vN
uN·v1 uN·v2 uN·vNhn uN·v1 uN·vN

Figure 2: Hard negative contrastive learning: Keyword substitution produces hard negative text samples, which are then randomly injected for each image u_i, replacing a simple negative sample in InfoNCE loss.

2 Vision-Language Representation Learning

Contrastive Learning. The objective of contrastive representation learning is to learn representations that are close to each other for similar samples and distant from each other for dissimilar samples. While many objectives originally addressed a single modality (Chopra et al., 2005; Schroff et al., 2015; Sohn, 2016; Oord et al., 2018), the idea can also be extended to multimodal training as well.

A successful example of multimodal learning is CLIP (Radford et al., 2021). CLIP is a Transformer-based model that consists of an image encoder and a text encoder, which are trained simultaneously. The objective is to maximize the cosine similarity of the image and text embeddings from the correct image-text pairs and to minimize the similarity between the incorrect pairs. A batch of N training samples (i.e., matching image-text pairs) results in a similarity matrix for each image-text combination. The main diagonal indicates the correct pair matches; the remaining entries correspond to negative entries. The symmetric cross-entropy loss is applied on $N \times N$ similarity scores.

This heuristical construction of negative samples has several issues. Negative captions can still match the given image in some cases, especially if the text is short and lacks details. Additionally, the negative pairs are often very dissimilar, which causes the model to decide only on coarse-grained features.

We illustrate this in Figure 1 with text-image pair I_1 and T_1^{pos}. The original learning process only uses negative text samples, such as T_2^{neg} and T_3^{neg}. Here, it is sufficient that the model can assign or negate objects from texts (i.e., nouns) to the objects in the image. The fact that no "person" and no "door" are present in image I_1 is sufficient for the model to discard these image descriptions. As a result, fine-grained concepts like object-color alignment, size, or spatial details ("cat under umbrella") are not necessary for reaching low loss. In this example, the model can rely solely on the presence and absence of specific objects. In this scenario, the caption can be seen as a bag of words without linguistic structure.

We address this problem by creating new hard negative data samples to learn more fine-grained concepts. See § 3 for details.

Related work. Several approaches incorporate hard negative samples in multimodal learning.

Radenovic et al. (2023) use importance sampling (based on Robinson et al., 2021), which up-samples hard negative samples and down-samples or ignores simple negatives. Yet, they reweight the simple negative samples per batch only but do not create new, challenging samples requiring fine-grained understanding.

Rösch and Libovický (2022) propose keyword permutation to create hard negative samples to learn spatial concepts. They added a spatial understanding classifier as an auxiliary pretraining objective and evaluated the models on visual question answering (VQA). Their model is based on LXMERT (Tan and Bansal, 2019) and not on a contrastive learning approach like CLIP.

Doveh et al. (2023) generate hard negatives using a rule-based procedure where they replace keywords. Moreover, they implement an approach where they randomly parse parts of speech and fill the mask using a BERT encoder with a plausible but wrong word. Unlike us, they do not incorporate hard negatives into the similarity matrix but use an auxiliary loss summed with the original contrastive loss. They use four distinct loss functions in total, introducing an additional layer of complexity to the overall procedure. In contrast, our approach uses the original loss function, with minimal modifications limited to the text inputs.

3 Contrastive Learning with Hard Negatives

We present a novel contrastive learning approach using sampled negative pairs and artificially generated textual hard negatives.

Instead of training with one positive sample and several weak negative samples, we create a scenario where models also minimize the similarity to hard negative *textual* samples. This forces the model to learn fine-grained concepts during training.

We use keyword substitutions for different concepts to break the correct meaning of an image caption. See Figure 1 for an example of the concept *color*. As a result, the new caption still lists, e.g., correct objects (e.g., "cat" and "umbrella") and actions (e.g., "cat sits") from the image but is no longer correct. We call this a "hard negative sample". Using these samples during training, we ensure that fine-grained concepts are learned. The idea to inject hard negative samples in contrastive learning is highlighted in Figure 2.

3.1 Creating Hard Negative Text Samples

For various concepts, we replace specific keywords using a *regex*-based tool. For example, we replace "white cat" with "brown cat" or "cat" with "dog". We create substitution heuristics for four different concepts, namely *colors*, *objects*, *size*, *location*:

- For *color*, any of the 9 most occurring color names in COCO can be replaced by any other color name.
- For *objects*, any of 80 object names can be replaced by any other. The 80 words originate from COCO object categories.
- For *location* keywords we use 12 one-to-one substitution relations.
- For *size* keywords we use 11 one-to-one substitution relations.

The full list of heuristics is shown in Table 3 in the Appendix. The created samples are used for training and evaluation in § 5.1 and § 5.3.

The keywords were selected based on dataset statistics. For specific applications, a domain expert may have to select other terms (e.g. specific colors in the fashion industry).

3.2 Training Details

In our experiments, we use CLIP's image and text encoder from the ViT-B/32 model,[1] which has 87 million and 63 million parameters respectively.

[1] https://huggingface.co/openai/clip-vit-base-patch32

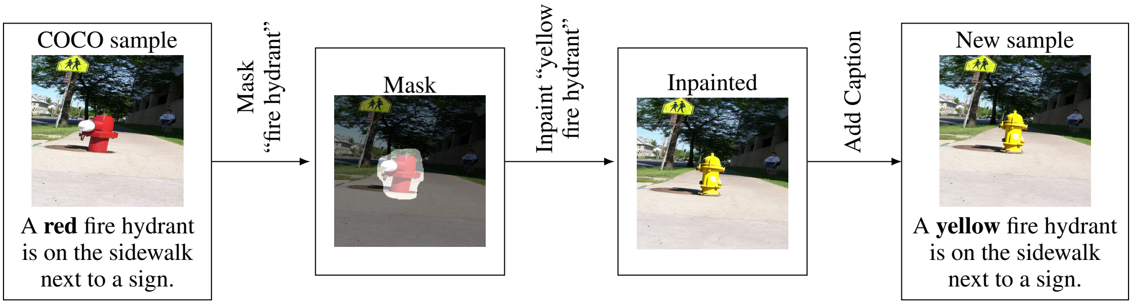

COCO sample
A **red** fire hydrant is on the sidewalk next to a sign.
Mask "fire hydrant"
Mask
Inpaint "yellow fire hydrant"
Inpainted
Add Caption
New sample
A **yellow** fire hydrant is on the sidewalk next to a sign.

Figure 3: Create hard negative image samples using open vocabulary segmentation for the masking prompt and text-to-image generation for the inpainting prompt. Additionally, a new correct caption is created manually. The InpaintCOCO dataset was created for concepts *object*, *color*, and *size*.

And hence, is a relatively small model compared with current multimodal GenAI models. Our code is modular, and any image and text encoder from `transformers` (Wolf et al., 2020) can be used in the framework.

We train all models using a batch size of 64. For concepts with multiple negative examples (i.e., *color*, *object*), we train models with 1, 2, and 3 hard negatives. This results in a proportion of hard negatives of 1.5%, 3%, and 4.5%, respectively. We only use one hard negative for the other concepts.

In all experiments, we use Adam optimizer (Kingma and Ba, 2014) with a weight decay of 0.1. We run evaluations with different learning rates (5×10^{-7}, 1×10^{-6}, 5×10^{-6}, 1×10^{-5}, 5×10^{-5}) and finally use 5×10^{-6}, since it leads to the best trade-off results for evaluation displayed in Figure 4. We do not use a learning rate scheduler since weights are already aligned. (This is not the case if encoders that were not trained simultaneously are used.) Checkpoints are saved every 10% of the data, and we train for 3 epochs.

The training time for all models was less than two hours on an Nvidia V100. Utilizing FP16 training, the GPU memory consumption remained below 12 GB.

3.3 Training Data

CLIP is pretrained on 400 million image-text pairs from web sources, and we continue the model pretraining. We use the COCO image captioning dataset (Lin et al., 2015), which has 591 thousand image-text pairs (2017 train version). For each concept, we filtered out samples where at least one keyword was present so that a keyword substitution could be applied. For evaluation, we use the validation set of COCO 2017. The respective dataset sizes are shown in the Appendix in Table 5.

4 InpaintCOCO: Challenge Set from the Visual Perspective

Many multimodal tasks, such as VL Retrieval and Visual Question Answering, present results in terms of overall performance. Unfortunately, this approach overlooks more nuanced concepts, leaving us unaware of which specific concepts contribute to the success of current models and which are ignored. More recent benchmarks attempt to assess particular aspects of vision-language models in response to this limitation. Some existing datasets focus on linguistic concepts utilizing one image paired with multiple captions; others adopt a visual or cross-modal perspective. In this study, we are particularly interested in fine-grained visual concept understanding, which we believe is not covered in existing benchmarks in sufficient isolation. Therefore, we create the InpaintCOCO dataset with image pairs with minimum differences that lead to changes in the captions.

Related Work. Benchmarks such as ARO (Yuksekgonul et al., 2023) or VL-CheckList (Zhao et al., 2022) evaluate models from the language perspective. ARO examines understanding of attributes and relations without using specific concepts such as color or size. VL-CheckList[2] is a dataset that investigates concrete concepts such as location, size, material, color, and relations.

On the other side, SVO (Hendricks and Nematzadeh, 2021) is a dataset that allows analysis from the visual perspective (2 images with 1 caption). Here, relations that deal with verbs are examined. To our knowledge, the only dataset that deals with fine-grained comprehension from a cross-model perspective is Winoground (Thrush et al., 2022). The dataset consists of two image-

[2] Parts of the dataset are not available anymore.

text pairs that are very similar to each other. This benchmark probes object relations that do not refer to specific concepts. The images show similar concepts but are very dissimilar in overall appearance. For samples of the two latter datasets, see Figure 5 in the Appendix. Both datasets contain real-world images that are in some ways similar in terms of objects, but the scenes still differ significantly. Therefore, it is difficult to tell which image differences cause the model predictions.

We overcome this limitation by creating InpaintCOCO, the first dataset with only minor changes in the *visual* components, so that concept comprehension can be analyzed in a more standardized setting.

Dataset Creation. The dataset creation process can be viewed a complement to textual hard-negative samples (§ 3) in the visual domain. Unlike keyword substitutions, this cannot be done automatically with sufficient accuracy. Even though image segmentation and generative inpainting tools reach impressive results, they still require human supervision to produce high-quality images. Creating a high-quality test set, therefore, requires annotation work.

To generate hard negative image samples, we need to change individual details in the image so that the textual image description no longer fits. The procedure is illustrated in Figure 3.

The annotation proceeds as follows: The annotators are provided with an image and decide if they want to edit it. If yes, they input the prompt for the object that should be replaced. Using the open vocabulary segmentation model CLIPSeg[3] (Lüddecke and Ecker, 2022) we obtain a mask for our object of interest (i.e., "fire hydrant"). Then, the annotator inputs a prompt for Stable Diffusion v2 Inpainting[4] (Rombach et al., 2022) (e.g. with the prompt "yellow fire hydrant"), which shows three candidate images. The annotators can try new prompts or skip the current image if the result is insufficient. Finally, the annotator enters a new caption that matches the edited image. See Appendix A for all details. The images and captions come from the COCO 2017 validation data. We only use images that contain the desired concept and where licenses allow adaptations.

We provide 452 images for the concept *object*, 465 for *color*, and 343 for *size*. In contrast to the training process, objects in images are only replaced with objects from the same COCO super category, i.e., "cat" with another animal or "chair" for another piece of furniture. Since *location* would require erasing at one spot and implanting objects at another in a nontrivial way (especially regarding depth), we discard this one concept in the newly created dataset. The dataset will be available upon publishing via the HuggingFace hub.

5 Experiments

We run several experiments to evaluate the proposed method. We compare the original OpenAI model (**Orig.**), with continued pretraining CLIP model using the classical contrastive learning approach (**Clas.**) and our method with 1 up to 3 hard negative values per batch (**HN1**, **HN2**, **HN3**). We measure both how well concepts are learned and whether the general image retrieval capability of the model changes on the COCO dataset (§ 5.1). Additionally, we evaluate the method using our InpaintCOCO challenge set (§ 5.2), and several other datasets (§ 5.3).

5.1 Fine-grained vs. Coarse Understanding

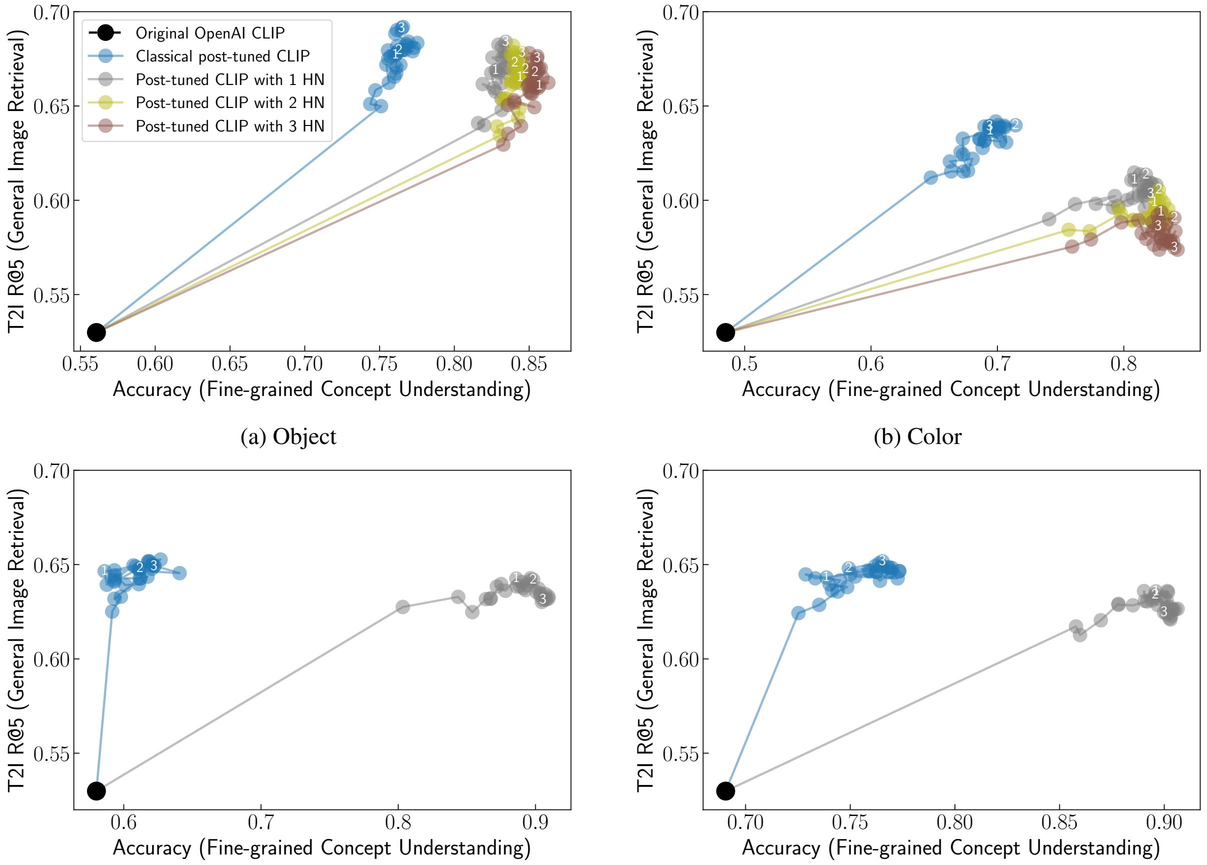
Fine-grained Concept Understanding. Here, we are interested in whether models have learned detailed concept knowledge. We pose the evaluation as a ranking problem with one image on one side and n different texts on the other side. Besides the correct text (containing the correct keyword), we generate all possible $n-1$ negative examples using the same procedure as in § 3.1 and rank the texts with the model. We evaluate the ranking using the top-1 accuracy. See Table 4 in the Appendix for some examples.

General Image Retrieval. We evaluate the general capabilities of our models using the COCO dataset. We want the general retrieval capability to remain high even though we train our models with a focus on one concept. We report text-to-image retrieval Recall@5 on the whole COCO validation set.

Results. Results are shown in Figure 4 (and exact results per epoch are displayed in Table 7 in the Appendix). The performance of the original OpenAI CLIP is shown with a black dot and constantly reaches the worst results. Continued pretraining on COCO massively increases the general retrieval

[3] https://huggingface.co/CIDAS/clipseg-rd64-refined

[4] https://huggingface.co/stabilityai/stable-diffusion-2-inpainting



(a) Object

(b) Color

(c) Location

(d) Size

Figure 4: Fine-grained Concept Understanding vs. General Image Retrieval: Results for four different concepts trained on corresponding dataset subsets. Checkpoints are evaluated after every 10% of the data (circles); checkpoints at epoch ends are marked with the respective numbers. The results are also in a table form in Table 7 in the Appendix.

performance for all concepts, showing the successful domain adaptation regarding this dataset. It also improves the concept understanding for all concepts except for *location*. We are especially interested in the trade-off between general retrieval and concept understanding for the different types of further pretrained models. The following values are based on model checkpoints at the end of the epoch.

For ***objects***, we observe a performance increase regarding the fine-grained comprehension from 0.56 for the original OpenAI CLIP models to 0.76 when further pretraining on COCO. Using the hard negatives approach, performance increases by 7 to 10 percentage points, depending on the hard negative proportion and training duration. Here, the general retrieval performance only drops by 1 to 2 percentage points in relation to the classical approach.

We observe a similar pattern can be seen for the concept ***color***. Understanding of the concept improved by 11 to 15 percentage points compared to the original training process. On the other hand, the general comprehension pattern loses 3 to 4 percentage points with HN1 and slightly more with hard negative values. In this case, one hard example seems sufficient to learn concepts.

A single hard negative sample per batch is enough to enhance spatial understanding (***location***). Here, concept comprehension improves by 29 or 30 percentage points (depending on the duration of training) to around 90%. Both the original CLIP model and the further pretrained model only achieved around 60%. The general understanding decreases by 1 percentage point only or by 2 percentage points with long training. A substantial improvement in location comprehension is achieved with negligible loss in overall understanding

We observe a similar pattern for the ***size*** concept. Further pretraining on COCO improves concept understanding by 5 to 8 percentage points to around 75%. On the other hand, if the new

Epoch	Orig.	Clas.	HN1	HN2	HN3
1	0.78	0.85	0.88	0.88	0.88
2	0.78	0.84	0.86	0.88	0.88
3	0.78	0.87	0.88	0.89	0.90

(a) Object

Epoch	Orig.	Clas.	HN1	HN2	HN3
1	0.62	0.83	0.89	0.89	0.89
2	0.62	0.83	0.90	0.90	0.91
3	0.62	0.83	0.90	0.91	0.92

(b) Color

Epoch	Orig.	Clas.	HN1
1	0.26	0.27	0.55
2	0.26	0.28	0.57
3	0.26	0.30	0.60

(c) Size

Table 1: Accuracy for fine-grained understanding from the visual perspective following Equation (1) for InpaintCOCO dataset.

learning concept is used, the result is 90%, which corresponds to an additional improvement of 15 percentage points. Depending on training time, the general image retrieval capabilities lose no or only 3 percentage points. As with location, an almost continuous improvement can be observed.

Across all concepts, hard negative contrastive learning can significantly increase concept understanding. This also applies to settings with just one hard example. It is shown that with a small adjustment in the objective, the models can learn a much more complex understanding of image and text. Meanwhile, the general ability to represent images and texts is hardly affected.

5.2 Challenge Set Results

The ranking-based evaluation in the previous section only assessed the model capabilities from the language perspective in a very similar setup to how the model was trained. This section assesses the model from the vision perspective using our InpaintCOCO challenge set.

We consider an image pair from the InpaintCOCO dataset correctly classified if the correct images would be more likely to be retrieved based on the original caption and the newly created caption. This leads to the formula,

$$\begin{aligned} \mathrm{sim}(i_{\mathrm{COCO}}, t_{\mathrm{COCO}}) > \mathrm{sim}(i_{\mathrm{inp}}, t_{\mathrm{COCO}}) \quad \wedge \\ \mathrm{sim}(i_{\mathrm{inp}}, t_{\mathrm{inp}}) > \mathrm{sim}(i_{\mathrm{COCO}}, t_{\mathrm{inp}}) \end{aligned} \tag{1}$$

with image i and text t, originating from the original COCO and InpaintCOCO dataset. The corresponding results for each concept are displayed in Table 1.

The results show that continued pretraining improves understanding of all three concepts (object, color, size). The improvements between the original OpenAI CLIP model and the further pretrained model are 6 to 9 percentage points for the object and 21 for the color concept. The improvements are less distinct for size, with a 1 to 4 percentage point gain, which aligns with the textual comprehension results displayed in Figure 4d.

For the ***object*** concept, hard negative training brought a 7 to 10 percentage point improvement for the textual viewpoint (see Table 7). For the evaluation from the visual perspective, improvements are 2 to 4 percentage points. This relatively smaller improvement is likely because each object could be replaced with the 79 other object names during training. Still in this evaluation, a replacement was only executed within the COCO super-category (5 to 10 object names per super-category). The differences in performance between the three models using hard negative training is $\leq$ 2 percentage points.

The evaluation of the ***color*** concept shows an improvement of 6 to 9 percentage points for the visual perspective. From the textual perspective (Table 7), the improvement was at over 11 percentage points. As before, using more hard negative samples during training does not further improve the performance systematically ($\leq$ 2 percentage point).

For the ***size*** concept, we see a big improvement for both perspectives when using hard negative models. From the visual perspective, there is an improvement of over 28 percentage points (Table 1c), and from the textual perspective, 13 to 16 percentage points.

The results show that training for just one epoch is sufficient for learning the concepts *object*, *color*, and *size*, and further training does not continue improving the results systematically. Additionally, using a single hard negative is sufficient to improve understanding of the concepts.

Dataset	Concept	Count	Orig.	Clas.	HN1	HN2	HN3
Flickr30k	object	30,926	.48	.66	.75	.77	.79
	color	45,003	.50	.64	.72	.73	.74
	location	20,097	.62	.64	.89		
	size	14,853	.75	.79	.92		
SBU	object	200,310	.33	.41	.48	.49	.50
	color	162,652	.51	.54	.61	.62	.62
	location	159,673	.61	.60	.79		
	size	47,069	.63	.65	.73		
Fashion200K	object	3,628	.24	.25	.35	.38	.39
	color	141,413	.68	.68	.69	.69	.70
NASA Earth Instagram	color	132	.43	.48	.55	.55	.58
Old Book Illustrations	object	124	.27	.35	.38	.35	.31
	location	95	.61	.50	.77		
	size	82	.61	.63	.79		

Table 2: Fine-grained concept understanding results (accuracy) for a diverse selection of datasets where the sample size is larger 50. Evaluated on dataset subsets where corresponding keywords are present.

5.3 Evaluations on other Datasets

We further investigate the performance of our model using more VL datasets. First, we evaluate the models using general VL datasets. The investigated concepts occur with different frequencies, and for high-quality results, it is important that these concepts are understood to increase the overall performance. Therefore, we further investigate fine-grained concept understanding on the Flickr30k (Young et al., 2014) and SBU Captioned Photo (Ordonez et al., 2011) datasets.

Fine-grained concept understanding is also important in specific domains. For example, in fashion, a correct assignment of garments and colors is important, not the mere presence of colors in the image. For this analysis, we evaluated our models on the very specific datasets Fashion200K (Han et al., 2017), NASA Earth Instagram,[5] and Old Book Illustrations.[6] These datasets are very heterogeneous in their appearance.

All models achieve good results except the *color* concept on the Fashion200k dataset and *object* concept on Old Book Illustrations. For the former, this is because images usually show garments with a distinct color. Yet, there is little background noise or noise from irrelevant items, which can confuse the color alignment of the model in this dataset. The latter shows old-fashioned drawings with objects very dissimilar to those in the COCO dataset. Our approach to learn concepts works very well for the remaining evaluations.

6 Conclusion

We introduce a robust method for enhancing fine-grained concept understanding with minimal impact on general retrieval capabilities using hard negative sampling in contrastive learning. We show that various concepts can be learned efficiently with minor text input adjustments. Moreover, improvements in concept understanding are observable after continued pretraining on only 10% of our data. Furthermore, one hard negative sample per image in a batch of 64 proves sufficient to incorporate the concept of interest into the model.

We comprehensively evaluate our method on several datasets, including our new challenge set. Our method outperforms classical contrastive learning on all investigated concepts. Existing datasets often focus on linguistic perturbations or use dissimilar images, precluding a structured evaluation of permuted visual concepts in isolation. To address this gap, InpaintCOCO represents the first dataset adjusting minor image parts in a controlled setting, facilitating cross-model fine-grained understanding. This ensures that the model's output is influenced only by one object and not the rest of the scene.

The results show that fine-grained concept understanding also generalizes to images of different styles when using InpaintCOCO and domain-specific datasets. Our method is data-efficient and requires only a little domain knowledge to design the hard negatives. This makes it particularly suitable for domain adaptation in image retrieval, as well as for developing new CLIP-based models, e.g., for object detection (Minderer et al., 2023).

[5] https://huggingface.co/datasets/nkasmanoff/nasa_earth_instagram

[6] https://huggingface.co/datasets/gigant/oldbookillustrations

Acknowledgements

The authors gratefully acknowledge the compute time granted by Monacum One at Bundeswehr University Munich. The work at CUNI, including annotation payments, was supported by Charles University project PRIMUS/23/SCI/023.

Limitations and Risks

Our research introduces a novel method for training CLIP aimed at incorporating concepts and representations that are challenging to learn using current approaches. In our experiments, we only used one specific CLIP model; however, we believe there is no reason why the method should not work or work systematically differently for smaller or larger CLIP models.

We conducted training with the well-studied COCO captioning dataset, which is standard in multimodal research. The proposed method is expected to show consistent performance also using other multimodal training datasets. Notably, evaluations on out-of-domain datasets, where the model was not trained, emphasize the robustness of our approach. One essential prerequisite for our methodology to work is the presence of keywords of interest in the training corpus and language and domain knowledge to decide how the keywords should be replaced. The keyword substitution will be more difficult in languages with more complex morphology than in English. Experiments involved using four concepts with a carefully chosen set of keywords. Depending on domain-specific tasks, other keywords might be of interest (e.g., a large list of garments for the fashion domain).

COCO is a dataset with image-text pairs where the captions are proper sentences, displaying a specific level of detail, and are carefully created by annotators. The images in the COCO dataset come from Flickr from 2014; therefore, they reflect the Flickr user structure at that time, i.e., the images mostly show the Western world and/or other countries from the Western perspective. The captions are in English. Thus, the model we developed does not generalize well beyond the Western world. However, we believe that is the limitation of the dataset, and the presented method itself is dataset agnostic.

The primary application goal of the models we worked with is to make image collections better accessible. Similar to other work on this VL modeling that enables better image understanding at scale, there is a risk of using technology based on the models for activities such as large-scale video surveillance.

References

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 539–546 vol. 1.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. 2023. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. Scaling open-vocabulary object detection. *Preprint*, arXiv:2306.09683.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. *arXiv preprint arXiv:2301.02280*.

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. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Philipp J. Rösch and Jindřich Libovický. 2022. Probing the role of positional information in vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1031–1041, Seattle, United States. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

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

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37, Abu Dhabi, UAE. Association for Computational Linguistics.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2023. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

A Appendix

Creating Hard Negative Samples In Table 3, we specify all used substitution keywords for the concepts *object*, *color*, *location*, and *size*.

Table 4 lists text samples for the fine-grained concept understanding task, which is used in § 5.1 for each concept.

Concept	Keywords
object	[person, bicycle, car, motorbike, aeroplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, sofa, potted plant, bed, dining table, toilet, tv monitor, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier, toothbrush]
color	[blue, red, green, yellow, black, white, brown, gray, orange]
location	left ↔ right, above ↔ below, under ↔ over, foreground ↔ background, in front of ↔ behind, back ↔ front
size	large ↔ small, little ↔ big, tall ↔ short, long → short, thin ↔ fat, huge ↔ tiny, giant → tiny

Table 3: All keywords that can be replaced for the four concepts. That are 80 for *object*, 9 for *color*, 12 for *location* and 11 for *size*. In lists, each word can be replaced by any other word. "↔" and "→" denote words that can be replaced in both and one direction, respectively.

COCO training and evaluation data. Table 5 shows the size of the training datasets and fine-grained concept understanding evaluation datasets. The images predominantly depict scenes from the USA and Western countries, and all captions are exclusively in English.

There are many large-scale VL datasets available to pretrain or further pretrain models (e.g., Conceptual Caption (Sharma et al., 2018) or LAION-5B (Schuhmann et al., 2022) with 3.3 million and 5 billion image-text pairs respectively). Yet, results in Figure 4 indicate that training on a small-sized COCO dataset (even for just one epoch) is sufficient to learn the concepts of interest.

Fine-grained VL Benchmarks. In Table 6, we list different benchmark datasets for fine-grained understanding in VL. Some image-text samples for ARO and Winoground are displayed in Figure 5 since these datasets provide two images per sample – similar to InpaintCOCO. However, unlike InpaintCOCO, the visual representations are very different regarding the scenes presented in these datasets.

✗	A small yellow person on a branch of a tree.
✗	A small yellow bicycle on a branch of a tree.
✗	...
✓	A small yellow **bird** on a branch of a tree.
✗	A small yellow cat on a branch of a tree.
✗	...

(a) Object (for all 80 object names)

✗	a blue cat sits under a black open umbrella.
✗	a red cat sits under a black open umbrella.
✗	a green cat sits under a black open umbrella.
✗	a yellow cat sits under a black open umbrella.
✗	a black cat sits under a black open umbrella.
✓	a **white** cat sits under a black open umbrella.
✗	a brown cat sits under a black open umbrella.
✗	a gray cat sits under a black open umbrella.
✗	a orange cat sits under a black open umbrella.

(b) Color

✓	a white cat sits **under** a black open umbrella.
✗	a white cat sits over a black open umbrella.

(c) Location

✓	A **small** yellow bird on a branch of a tree.
✗	A large yellow bird on a branch of a tree.

(d) Size

Table 4: Exemplary image descriptions that are used as text samples in the fine-grained concept understanding task (see § 5.1) for the example images displayed in Figure 1 or Figure 3.

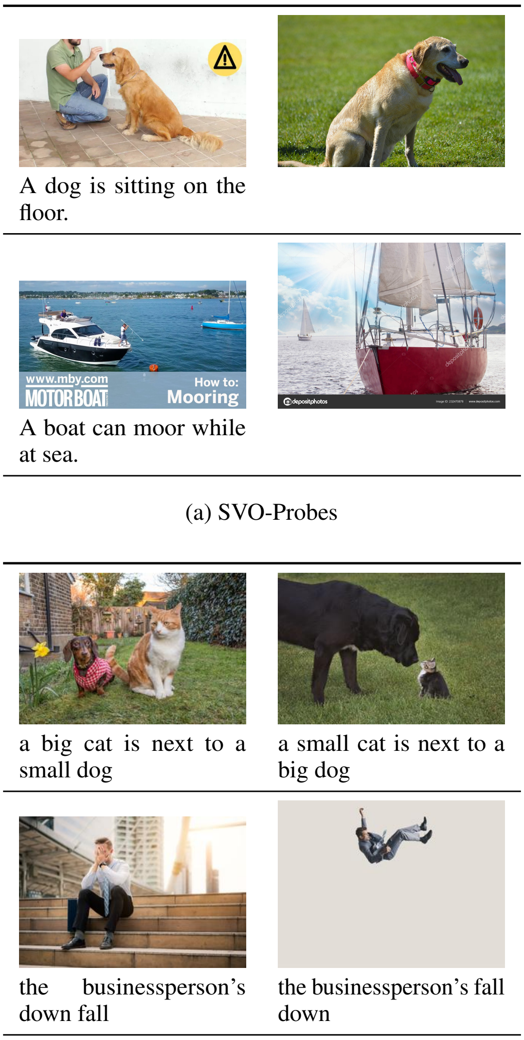
Concept	Further pretraining dataset size	Evaluation dataset size
object	305,056	12,907
color	92,006	3,907
location	58,156	2,527
size	60,626	2,601

Table 5: COCO dataset size for all concepts used for further pretraining CLIP and evaluation.

Challenge Set Creation. Undergraduate student workers created the challenge set. They were provided with an interactive Python environment with which they interacted via various prompts and inputs. The description of the task and the problem of the research question was made available to them (see Figure 6). In addition to a detailed written explanation of how the tool works, they were also

Dataset	Size	Perspective	Samples	Tasks
ARO	77k	Linguistic	1 image ↔ 2 texts	Attributes, Relations, Order understanding
VL-CheckList	410k	Linguistic	1 image ↔ 2 texts	Attributes, Relations, Object understanding
SVO-Probs	48k	Visual	2 images ↔ 1 text	Relations (Verb Understanding)
Winoground	400	Cross-modal	2 images ↔ 2 texts	Relations
InpaintCOCO	1,260	Cross-modal	2 images ↔ 2 texts	Attributes, Object understanding

Table 6: Datasets for fine-grained understanding in VL.


A dog is sitting on the floor.
www.mby.com
MOTOR BOAT
How to: Mooring
depositphotos
A boat can moor while at sea.
a big cat is next to a small dog
a small cat is next to a big dog
the businessperson's down fall
the businessperson's fall down

(a) SVO-Probes

(b) Winoground

Figure 5: Samples from visual and cross-model datasets.

given "best practices," which were created by one student and reviewed by the authors.

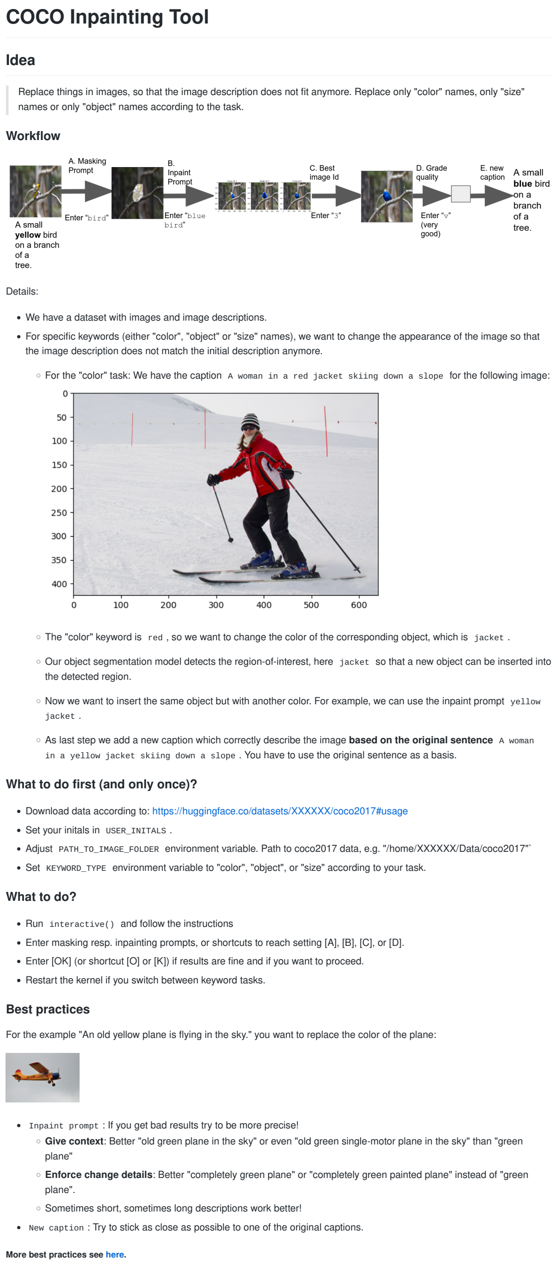
For color, any other *color* and for *size*, the opposite statement can be chosen. Yet, within *objects*, the students were asked to replace with objects from the same COCO super-category (to ensure that no "plain" needs to be inpainted in an indoor scene). There are 12 super categories for the 80 object names: person, vehicle, outdoor, animal, accessory, sports, kitchen, food, furniture, electronic, appliance, and indoor.

The workflow comprises these steps:

1. A random COCO (2017 validation) image is shown with all its captions containing a concept keyword.
2. The annotator enters a masking prompt for the segmentation task based on the object of interest (e.g., "fire hydrant" in Figure 3). They can also enlarge the mask within the x and y dimensions by passing additional parameters. This is useful if a larger object is to be inserted into the image. Several attempts can be made until the mask meets the requirements. Only then the next step is carried out.
3. Then, the annotator enters an inpainting prompt (the image generation takes roughly 1 minute). They are provided with three different inpainted images. They proceed if at least one high-quality image has been generated.
4. The best image is chosen from the three proposals.
5. Based on the selection before, they rate the pictures as "very good" or "okay".
6. Finally, a new, correct caption is added based on one of the original COCO captions.

A subset of students had pre-existing roles within the university, while others were purposefully recruited for the designated task. The compensation for student assistants adhered to the legally stipulated wages in their respective countries, amounting to CZK 300.00 per hour in the Czech Republic and EUR 12.00 per hour in Germany.

Challenge Set Details. Our InpaintCOCO challenge set is based on the famous COCO dataset. All captions follow "Creative Commons Attribution 4.0 License" and hence can be changed. Images originate from Flickr, and have diverse licenses ("Attribution License", "Attribution-NonCommercial-ShareAlike License", "Attribution-NonCommercial License", "Attribution-ShareAlike License") which all allow scientific usage and modification (like inpaining).



COCO Inpainting Tool

Idea

> Replace things in images, so that the image description does not fit anymore. Replace only "color" names, only "size" names or only "object" names according to the task.

Workflow

Details:

- We have a dataset with images and image descriptions.
- For specific keywords (either "color", "object" or "size" names), we want to change the appearance of the image so that the image description does not match the initial description anymore.
  - For the "color" task: We have the caption `A woman in a red jacket skiing down a slope` for the following image:

  - The "color" keyword is `red` , so we want to change the color of the corresponding object, which is `jacket` .
  - Our object segmentation model detects the region-of-interest, here `jacket` so that a new object can be inserted into the detected region.
  - Now we want to insert the same object but with another color. For example, we can use the inpaint prompt `yellow jacket` .
  - As last step we add a new caption which correctly describe the image **based on the original sentence** `A woman in a yellow jacket skiing down a slope` . You have to use the original sentence as a basis.

What to do first (and only once)?

- Download data according to: https://huggingface.co/datasets/XXXXXX/coco2017#usage
- Set your initals in `USER_INITALS` .
- Adjust `PATH_TO_IMAGE_FOLDER` environment variable. Path to coco2017 data, e.g. "/home/XXXXXX/Data/coco2017"`
- Set `KEYWORD_TYPE` environment variable to "color", "object", or "size" according to your task.

What to do?

- Run `interactive()` and follow the instructions
- Enter masking resp. inpainting prompts, or shortcuts to reach setting [A], [B], [C], or [D].
- Enter [OK] (or shortcut [O] or [K]) if results are fine and if you want to proceed.
- Restart the kernel if you switch between keyword tasks.

Best practices

For the example "An old yellow plane is flying in the sky." you want to replace the color of the plane:

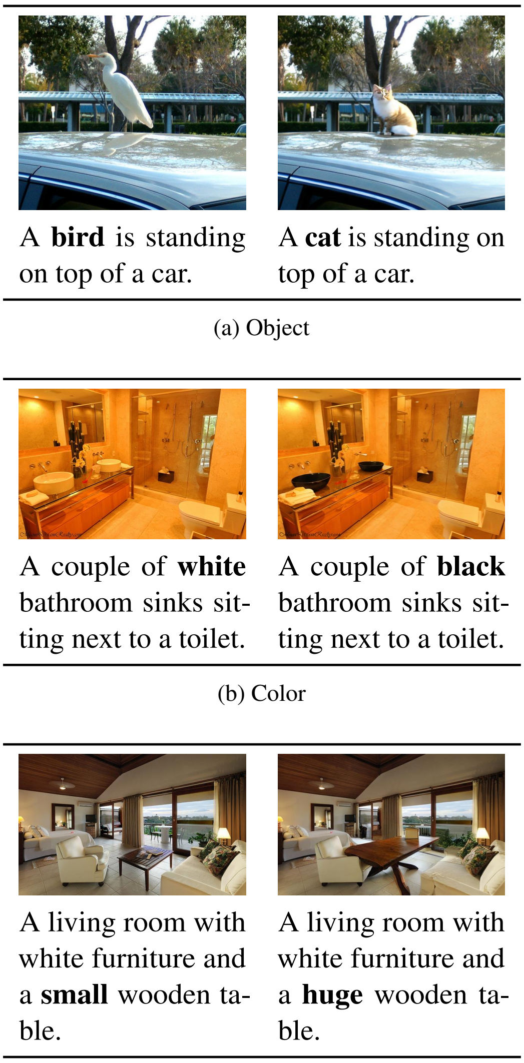
- `Inpaint prompt` : If you get bad results try to be more precise!
  - **Give context**: Better "old green plane in the sky" or even "old green single-motor plane in the sky" than "green plane"
  - **Enforce change details**: Better "completely green plane" or "completely green painted plane" instead of "green plane".
  - Sometimes short, sometimes long descriptions work better!
- `New caption` : Try to stick as close as possible to one of the original captions.

More best practices see here.

Figure 6: Instructions of inpainting tool provided to student workers.

Individual licenses are listed in each sample of the dataset.

Experiment Results. For numeric results from Figure 4 see Table 7. In this Table fine-grained concept understanding results (Accuracy) and COCO text-to-image retrieval results (T2I R@5) are presented. "Orig.", "Clas.", "HN1", "HN2", and "HN3" indicate the original OpenAI CLIP model, the classical further pretrained model, and models trained with 1, 2, or 3 hard negative samples.



A **bird** is standing on top of a car.

A **cat** is standing on top of a car.

(a) Object

A couple of **white** bathroom sinks sitting next to a toilet.

A couple of **black** bathroom sinks sitting next to a toilet.

(b) Color

A living room with white furniture and a **small** wooden table.

A living room with white furniture and a **huge** wooden table.

(c) Size

Figure 7: InpaintCOCO samples for all concepts.

		Accuracy					T2I R@5				
Concept	Epoch	Orig.	Clas.	HN1	HN2	HN3	Orig.	Clas.	HN1	HN2	HN3
object	1	.56	.76	.83	.84	.86	.53	.68	.67	.67	.66
	2	.56	.76	.84	.85	.85	.53	.68	.67	.67	.67
	3	.56	.76	.83	.85	.86	.53	.69	.68	.68	.68
color	1	.48	.69	.81	.82	.83	.53	.64	.61	.60	.59
	2	.48	.71	.82	.83	.84	.53	.64	.61	.61	.59
	3	.48	.69	.82	.83	.84	.53	.64	.60	.59	.57
location	1	.58	.59	.89			.53	.65	.64		
	2	.58	.61	.90			.53	.65	.64		
	3	.58	.62	.91			.53	.65	.63		
size	1	.69	.74	.90			.53	.64	.64		
	2	.69	.75	.90			.53	.65	.63		
	3	.69	.77	.90			.53	.65	.62		

Table 7: Accuracy of fine-grained concept understanding (evaluated on dataset subsets) and COCO text-to-image Recall@5 for general image retrieval (evaluated on whole dataset) for models trained on respectively concepts. Checkpoints for epochs 1 to 3.