RCA-NOC: Relative Contrastive Alignment for Novel Object Captioning

Anonymous ICCV submission

Paper ID 5814

Abstract

In this paper, we introduce a novel approach to novel object captioning which employs relative contrastive learning to learn visual and semantic alignment. Our approach maximizes compatibility between regions and object tags in a contrastive manner. To set up a proper contrastive learning objective, for each image, we augment tags by leveraging the relative nature of positive and negative pairs obtained from foundation models such as CLIP. We then use the rank of each augmented tag in a list as a relative relevance label to contrast each top-ranked tag with a set of lower-ranked tags. This learning objective encourages the top-ranked tags to be more compatible with their image and text context than lower-ranked tags, thus improving the discriminative ability of the learned multi-modality representation. We evaluate our approach on two datasets and show that our proposed RCA-NOC approach outperforms state-of-the-art methods by a large margin, demonstrating its effectiveness in improving vision-language representation for novel object captioning.

1. Introduction

Describing novel objects unseen in training data is a highly desired capability for a real-world image captioning model. Conventional image captioning models [4, 28, 20] often fail to describe novel objects because they only cover limited visual concepts and generalize poorly to images in the wild [26]. To overcome this limitation, approaches that rely on object detection as the external resource[27, 20, 33, 8, 25, 7, 18, 13, 35] have been widely explored and demonstrated breakthroughs in vision-language (VL) understanding.

Although object detection models (*e.g.*, Faster RCNN [23]) have been improved to recognize a wide range of objects including novel ones like zero-shot object detection [14], using object detection in novel captioning models brings a new challenge. VIVO [13] leverages extra object tags to pre-train a visual vocabulary and help image captioning generalize to new categories. NOC-REK [29] tries to augment object tags in the training stage based on the similarity between region and word. Such methods simply concatenate regions, object tags, and caption features as input to a Transformer-based model and use masked token reconstruction to implicitly learn an alignment between vision and language. Although such an alignment can help bring regions and words of the same concept closer, it lacks an effective mechanism to push away irrelevant concepts. As a result, it still makes mistakes on some confusing concepts, as shown in Figure 1.

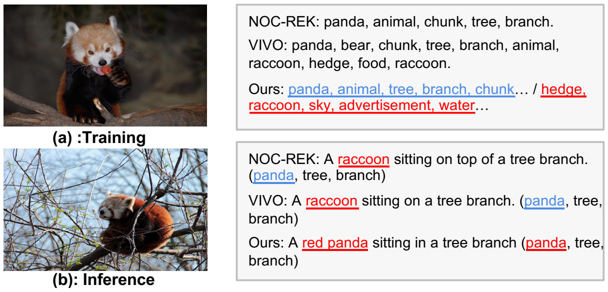

NOC-REK: panda, animal, chunk, tree, branch.
VIVO: panda, bear, chunk, tree, branch, animal, raccoon, hedge, food, raccoon.
Ours: panda, animal, tree, branch, chunk... / hedge, raccoon, sky, advertisement, water...
(a) :Training
NOC-REK: A raccoon sitting on top of a tree branch. (panda, tree, branch)
VIVO: A raccoon sitting on a tree branch. (panda, tree, branch)
Ours: A red panda sitting in a tree branch (panda, tree, branch)
(b): Inference

Figure 1: An illustrative example of our method to leverage relative semantic relevance to achieve modality alignment. Our method could give accurate captions `"A cute red panda sitting in a tree branch"` conditioned on the objects `"red panda"` while VINVL+VIVO generates a wrong caption `"A cute raccoon sitting in a tree branch"`. This inference result shows that our method can differentiate some confusing objects and generate accurate captions for novel objects when the object detection is well aligned with other modalities.

Different from previous methods, this study aims to learn visual and semantic alignment in a contrastive manner. Gupta *et al*. [10] showed that modality alignment could be achieved by maximizing the information lower bound between an image and its object tags. That is, given pairs of image and object tags, we can maximize the compatibility between tags and their attention-weighted region representations, compared to regions and non-corresponding tags. However, there are two critical problems that need to be further addressed: 1) how to effectively generate contrastive tags (augmented tags) that are closely relevant to an image, and 2) how to design a proper contrastive learning objective that allows the model to effectively leverage the contrastive tags to align vision and semantics.

To tackle the first problem, we utilize CLIP [22] to create a list of contrastive tags which are closely linked to an image and contain global structural information and high-level concepts describing scenes. This approach aids in the discovery of useful tags that are essential for contrastive learning. To address the second problem, a proper contrastive learning objective needs to be explored. The major challenge here is that the augmented contrastive tags are inaccurate and inevitably noisy. We cannot simply treat one tag as positive and others as negative to perform contrastive learning, because the augmented tags might be highly correlated or similar to each other. To tackle this problem, we leverage the relative relevance, rather than the absolute relevance, of the augmented tags, which is more robust to data noise.

Specifically, given an image, for each of its labeled object tags, we generate a ranked list of contrastive tags using CLIP. We regard the rank of each augmented tag as its relative semantic relevance with the image. In general, the top-ranked tags are assumed to be more relevant to the image than the lower-ranked tags. We divide the list into two parts: the (relatively) relevant part with higher rank and the (relatively) irrelevant part with lower rank (Figure 1 (a)). In our proposed objective, we treat each tag in the relevant part as positive and all tags in the irrelevant part as negative to perform contrastive learning. Note that we do not let tags both in the relevant part contrast with each other as they might be highly correlated concepts. In this way, our approach weakens the strict assumption of contrastive learning in previous works and exploits the relative ranking in a loose form to achieve modality alignment.

We conduct experiments on the Nocaps and Held-Out COCO datasets to demonstrate the effectiveness of RCA-NOC. Our contributions can be summarized as follows:

- We propose Relative Contrastive Alignment (RCA) to learn the relative semantic relevance in a loose form by maximizing the compatibility between regions and their relevant tags compared with regions and irrelevant tags to achieve vision and language alignment and improve the discriminative ability of the multi-modality representation.
- A method called Uncertainty-Aware Selection and Reweighting (UASR) is proposed to estimate and exploit the uncertainty of each contrastive sample to mitigate the negative effect brought by noisy tags. UASR can effectively prioritize highly reliable samples and demote false positives and false negatives.
- We validate the proposed method on the Nocaps and Held-Out COCO benchmarks, which outperforms other state-of-the-art methods by a large margin.

2. Related work

Novel object captioning aims to describe images with objects that are unseen in the training stage (we define these objects as novel objects) where many methods. [27, 20, 33, 8, 7, 18, 13, 35, 22, 31] have been proposed. Early works such as Hendricks *et al.* [12] and Venugopalan *et al.* [27] utilize unpaired labeled image and sentence data to enhance semantically visual concepts. Recent studies propose to explicitly leverage the object detection results for NOC, Lu *et al.* [20], Wu *et al.* [33], and Demirel *et al.* [8] fill the generated template sentence with objects detected by object/novel object detectors. Chen *et al.* [7] combine object detector and human attention to identify novel objects. In addition, Li *et al.* are the first to utilize semantics in VLP tasks, which are further extended by Zhang *et al.* [35]. Hu *et al.* [13] built upon [35] and propose to leverage extra region-tag pairs to conduct pretraining. Duc *et al.* [29] tried to augment object tags in the training stage based on the similarity of regions and objects.

However, most aforementioned methods for NOC ignore the misalignment problem of object tags, thereby failing to fully exploit the semantic relationship between vision and language (Figure 1), which we argue is crucial to the quality of generated captions. In this paper, we propose a simple but effective contrastive learning objective to learn the relative semantic relevance in a loose form where the object tags could be explicitly aligned with their corresponding image feature representations in a semantic space.

Contrastive Learning aims to learn discriminative representations to distinguish an image from others. Many methods [6, 11, 32, 16, 10, 22, 17] have shown their effectiveness. For example, Chen *et al.* [6] proposed to learn visual representations by maximizing agreement between differently augmented views of the same image via a contrastive loss. He *et al.* [11] proposed Momentum Contrast (MoCo) for unsupervised visual representation learning. Wei *et al.* [32] utilized contrastive learning in the detection task. Li *et al.* [16] focused on the problem of contextual outpaint-

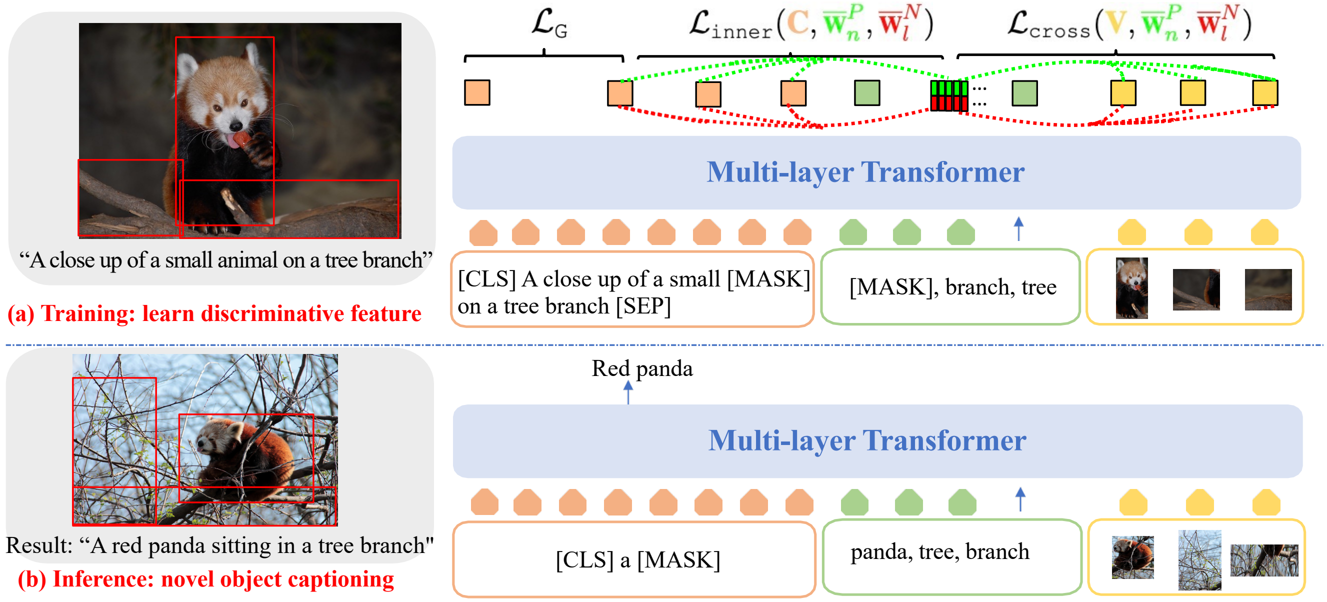

$\mathcal{L}_{\mathrm{G}}$
$\mathcal{L}_{\mathrm{inner}}(\mathbf{C}, \overline{\mathbf{w}}_n^P, \overline{\mathbf{w}}_l^N)$
$\mathcal{L}_{\mathrm{cross}}(\mathbf{V}, \overline{\mathbf{w}}_n^P, \overline{\mathbf{w}}_l^N)$
Multi-layer Transformer
"A close up of a small animal on a tree branch"
(a) Training: learn discriminative feature
[CLS] A close up of a small [MASK] on a tree branch [SEP]
[MASK], branch, tree
Red panda
Multi-layer Transformer
Result: "A red panda sitting in a tree branch"
(b) Inference: novel object captioning
[CLS] a [MASK]
panda, tree, branch

Figure 2: The main pipeline of our proposed RCA-NOC, including training and inference: In addition to the normal caption loss L_G in the training process, we also compute the cross-modality and inner-modality loss L_{cross} and L_{inner} in Eq. 15 and Eq. 16 by contrasting the positive tags against the negative tags. The corresponding relationship is denoted in green and red line separately. During inference, our model could accurately generate the caption that contains novel objects.

ing. Gupta *et al*. [10] proposed a method CPG[1] to find hard negative words by replacing a word in a caption.

Other data-driven studies such as CLIP [22] and ALIGN [17] focus on learning a corresponding relative relationship from massive web data. CLIP [22] predicts which text goes with which image and learns a relative image-text corresponding relationship from broad web data with noisy supervision. ALIGN [17] further scales up CLIP by leveraging a noisy dataset that covers more than one billion image-text pairs. These methods achieve remarkable results but require massive fully-annotated data, which are difficult to obtain.

3. Proposed Methods

We propose to enhance the modality alignment by explicitly injecting visual semantics. These semantics are extracted for each image, obtained with existing foundation models such as CLIP.

There are two motivations behind our approach. First, by maximizing the compatibility between regions and their relevant tags compared with regions and irrelevant tags, we could not only learn vision-semantic alignment and improve the discriminative ability of the multi-modality representation, Second, we introduce a relative contrastive learning objective that considers the relative relationships between positive and negative examples, rather than using absolute prototype-contrastive learning methods. This approach is more generalized and can be easily integrated into any existing NOC method.

3.1. Visual Semantics Extraction

Different from other approaches (e.g., [18, 13]) that extract object tags using pretrained object detectors, we use an off-the-shelf foundational model to extract more diverse, larger and semantically meaningful set of visual semantics. We show in section 4, that this approach helps to capture high level and global semantics describing the scenes, which are hard to obtain with other approaches (e.g. object detectors).

We use a pretrained CLIP (ViT-B/16) model to obtain the embeddings of all the images and the extracted semantics. For each image, we compute its cosine similarity with all the embedded semantics and select the top M similar semantics as augmented tags $T = \{t_l\}_{l=1}^M$. The augmented tags will be ranked using global (image-level) cosine similarity $p(t_l)$. Then we utilize the rank of each augmented tag in T as a relative relevance label. In general, the top-ranked tags are assumed to be more relevant to the image than the lower-ranked tags. We divide T into positive and negative tags: positive tags being the (relatively) relevant tags with higher rank $T^P = \{t_n^P\}_{n=1}^K = \{t_l\}_{l=1}^K$ and negative tags being the (relatively) irrelevant tags with

[1] For better illustration, we use the abbreviation CPG to denote the paper "Contrastive Learning for Weakly Supervised Phrase Grounding" [10]

lower rank $T^N = \{t_l^N\}_{l=1}^K = \{t_l\}_{l=K+1}^{2K}$ ($M = 2K$).

Here we propose a simple yet effective augmentation technique. Having a set of visual semantics extracted for each image, instead of considering all the visual semantics at once, we sample randomly a fraction of these augmented tags at each iteration step. Hence, we could not only prevent the model from overfitting on specific semantics and potentially disregard the image or other semantics during training, but also let the model see different combinations of visual semantics, which helps to have more diversity.

3.2. Relative Contrastive Alignment

InfoNCE [21] is a typical type of contrastive loss function used for self-supervised learning that contrasts one positive sample against a set of negative samples. CPG [10] builds upon InfoNCE and proposes a compatibility function for regions and a caption word. Both InfoNCE and CPG utilize the absolute relevance to optimize the contrastive loss.

However, such a contrastive learning objective is hard to optimize for object tags since they may be highly correlated to each other. In RCA-NOC, we focus on the relative semantic relevance and utilize the attention-based compatibility function to measure the relativity information for regions and a set of object tags, which is more robust to data noise.

Specifically, we first compute the dot product for a region-tag pair.

$$s_{jk} = \mathbf{w}_j^T \mathbf{v}_k / \sqrt{d}, \tag{1}$$

where $\mathbf{w}_j$ and $\mathbf{v}_k$ refer to the corresponding embeddings for tag and region, and d is the feature dimension. Here, s_{jk} represents the similarity between the j-th tag and the k-th region. To find a contextualized region representation for the j-th tag, we define $\mathbf{a}_j^v$ as follows.

$$\mathbf{a}_j^v = \sum_{k=1}^{K} \alpha_{jk} \mathbf{v}_k, \tag{2}$$

where
$$\alpha_{jk} = \frac{e^{s_{jk}}}{\sum_{k'=1}^{K} e^{s_{jk'}}}, \tag{3}$$

To measure the compatibility between a tag $\mathbf{w}_j$ and its contextualized region representation $\mathbf{a}_j^v$, we define

$$\phi(\mathbf{V}, \mathbf{w}_j) = \mathbf{w}_j^T \mathbf{a}_j^v. \tag{4}$$

In this way, we can derive our cross-modality contrastive loss function L_{cross}.

$$\mathcal{L}_{\text{cross}}(\mathbf{V}, \mathbf{W}^P, \mathbf{W}^N) = -\frac{1}{K}\sum_{n=1}^{K} \log\left(\frac{e^{\phi(\mathbf{V},\mathbf{w}_n^P)}}{e^{\phi(\mathbf{V},\mathbf{w}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{V},\mathbf{w}_l^N)}}\right), \tag{5}$$

where $\mathbf{W}^P = \{\mathbf{w}_l^P\}_{l=1}^K$ and $\mathbf{W}^N = \{\mathbf{w}_n^N\}_{n=1}^K$ are the word embeddings for positives/negatives T^P and T^N, extracted from CLIP. $\phi()$ is the compatibility function defined in Eq. 4.

This equation encourages each positive tag $\mathbf{w}_n^P$ to be more compatible with image regions $\mathbf{V}$ than all negative tags in $\mathbf{W}^N$. Essentially, if tag t_j is not relevant to the image (negative), its representation $\mathbf{w}_j$ should not be similar to its contextualized region representation $\mathbf{a}_j^v$ since it would not be able to collect good information while computing $\mathbf{a}_j^v$. Note that we do not let two positive tags (top-ranked tags) contrast with each other as it is hard to tell which one is more relevant.

Inner modality Alignment. To further enhance modality alignment, we also compute the inner-modality contrastive loss over tag-caption pairs. The formulation for inner-modality contrastive loss is similar to cross-modality contrastive loss but slightly different. We enforce contrastive learning not to be between tags and all caption words, but just between tags and noun caption words. This is because a caption may have many irrelevant tokens, such as ”the, a, is”, which will harm our alignment process and cause redundant computational costs.

Finally, we can derive our inner-modality contrastive loss function L_{inner}.

$$\phi(\mathbf{C}, \mathbf{w}_l) = \mathbf{w}_l^T \mathbf{a}_l^c, \tag{6}$$

$$\mathcal{L}_{\text{inner}}(\mathbf{C}, \mathbf{W}^P, \mathbf{W}^N) = -\frac{1}{K}\sum_{n=1}^{K} \log\left(\frac{e^{\phi(\mathbf{C},\mathbf{w}_n^P)}}{e^{\phi(\mathbf{C},\mathbf{w}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{C},\mathbf{w}_l^N)}}\right), \tag{7}$$

where this time we take noun caption words $\mathbf{C} = \{\mathbf{c}_i\}_{i=1}^P$ in Eq. 6 as input. By further considering inner-modality alignment, we not only learn caption-tag semantics to connect relevant caption tokens, but also enable region-caption interactions.

3.3. Uncertainty-Aware Selection and Re-weighting

The augmented contrastive tags are often noisy and would harm the discriminative ability and robustness of the contrastive learning process. Therefore, we design a sample selection strategy to deal with noisy contrastive tags (*e.g.*, false positives/negatives).

Uncertainty-Aware Selection. Specifically, we first use the local cosine similarity to calculate the correlation between a region and a tag and filter false positives/negatives. The stronger the correlation, the more reliable a positive tag is and the more uncertain a negative tag.

$$\mathrm{u}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\|\,\|\mathbf{w}\|}, \tag{8}$$

Then we retrieve L tags and their corresponding representations $\mathbf{H} = \{\mathbf{w}_j\}_{j=1}^L$ for the given L region features

$\{\mathbf{v}_i\}_{i=1}^{L}$. We use arg max to choose the most correlated tag t_j for region $\mathbf{v}_i$.

$$j = \arg\max_{k}(\mathrm{u}(\mathbf{v}_i, \mathbf{w}_k)), \tag{9}$$

According to the score in Eq. 8, the corresponding top-L tags will be listed as confusion samples for negatives (actually positive), and these confusion samples will be removed from the original contrastive tags. Similarly, we will also choose the true positives based on these top-L tags since tags with low-similarity scores should be regarded as outliers to the positives and they are too sparse to have a positive influence on the shape of embedding space. In this way, we can derive Eq. 10, where / is the removing set operation, $\cap$ is the intersection set operation. $\overline{\mathbf{W}}^P = \{\bar{\mathbf{w}}_l^P\}_{l=1}^K$ and $\overline{\mathbf{W}}^N = \{\bar{\mathbf{w}}_n^N\}_{n=1}^K$ is the final positive/negative tags' embedding sets. Note that if the set length of $\overline{\mathbf{W}}^P/\overline{\mathbf{W}}^N$ is less than M, we will over-sample $\overline{\mathbf{W}}^P/\overline{\mathbf{W}}^N$ until the set length is equal to M.

$$\overline{\mathbf{W}}^P = \mathbf{W}^P \cap \mathbf{H}, \overline{\mathbf{W}}^N = \mathbf{W}^N/\mathbf{H}, \tag{10}$$

Uncertainty-Aware Re-weighting. To further mitigate the negative effect of noisy tags, we use the correlation function defined above to enhance reliable samples and reduce the influence of high-uncertainty samples. The information from more reliable samples should have a larger impact on the shape of the embedding and vice versa. Specifically, we first introduce the weight in Eq. 11.

$$q_1(\bar{\mathbf{w}}_n^P) = \exp\left(\mathrm{u}(\mathbf{v}_k, \bar{\mathbf{w}}_n^P)\right), \tag{11}$$

$$j = \arg\max_{k}(\mathrm{u}(\mathbf{v}_i, \bar{\mathbf{w}}_k^P)), \tag{12}$$

$$q_2(\bar{\mathbf{w}}_n^P) = p(t_n), \tag{13}$$

where $q_1(\bar{\mathbf{w}}_n^P)$ in Eq. 11 is the uncertainty score computed by the most relevant region with the corresponding tag. Moreover, we further consider how certain each tag is and combine the corresponding score $q_2(\bar{\mathbf{w}}_n^P)$ in Eq. 13 with $q_1(\bar{\mathbf{w}}_n^P)$ to derive the final uncertainty score $q(\bar{\mathbf{w}}_n^P)$ in Eq. 14. Here $p(t_n)$ in Eq. 13 is the global similarity mentioned in Section 3.1.

$$q(\bar{\mathbf{w}}_n^P) = q_1(\bar{\mathbf{w}}_n^P) * q_2(\bar{\mathbf{w}}_n^P), \tag{14}$$

Thus, our method could mitigate the negative effect of noisy contrastive tags by considering the similarity from both the local region features (region-tag) and the global structural information (image-tag). Finally, we could rewrite our contrastive loss function in Eq. 15, Eq. 16.

$$\mathcal{L}_{\mathsf{cross}}(\mathbf{V}, \overline{\mathbf{W}}^P, \overline{\mathbf{W}}^N) = -\frac{1}{K}\sum_{n=1}^{K} -q(\bar{\mathbf{w}}_n^P)\log\left(\frac{e^{\phi(\mathbf{V},\bar{\mathbf{w}}_n^P)}}{e^{\phi(\mathbf{V},\bar{\mathbf{w}}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{V},\bar{\mathbf{w}}_l^N)}}\right), \tag{15}$$

$$\mathcal{L}_{\texttt{inner}}(\mathbf{C}, \overline{\mathbf{W}}^P, \overline{\mathbf{W}}^N) = -\frac{1}{K}\sum_{n=1}^{K} -q(\bar{\mathbf{w}}_n^P)\log\left(\frac{e^{\phi(\mathbf{C},\bar{\mathbf{w}}_n^P)}}{e^{\phi(\mathbf{C},\bar{\mathbf{w}}_n^P)} + \sum_{l=1}^{K} e^{\phi(\mathbf{C},\bar{\mathbf{w}}_l^N)}}\right), \tag{16}$$

4. Experiments

4.1. Experimental Setup

Datasets. Our main experiments and ablation studies are based on the Nocaps[2] dataset. Our method was built with PyTorch, and we used a pre-trained BERT-base model from Huggingfaces [1] for parameters initialization, and no ground-truth tags are used on the Nocaps validation and test sets. For the training stage, we use the COCO training set which consists of 118K images, each with 5 captions. We evaluate our model on the validation and test sets of the Nocaps dataset, which consist of 4.5K and 10.6K images from the Open Images validation and test sets, respectively. Additionally, we test the proposed method on the Held-Out COCO [12], which is a subset of MS COCO [19] where the following eight object categories are excluded from the training set: bottle, bus, couch, microwave, pizza, racket, suitcase, and zebra. We randomly split the COCO validation set and use half of it for validation and the other half for testing, each with 20,252 images. In addition, we empirically set $M = 50$ to generate augmented object tags for the best performance.
Implementation Details. In the training stage, the model is trained for 30 epochs with a batch size of 512 and a learning rate of 10^{-4}, optimized using the cross-entropy loss and our contrastive loss. We set the maximum caption length to 40 and the maximum tag length to 30. To further boost the performance, we also perform the SCST optimization (Rennie *et al.* [24]) with a learning rate of 1.4×10^{-6} for 10 epochs. During inference, we use greedy decoding to generate image captions with a maximum length of 20. Our model is trained with 8 A100 GPUS and takes 1 day to train.

4.2. Quantitative evaluation.

Results on Nocaps dataset. In this section, we extensively compare our frameworks with previous methods on the Nocaps benchmark. The other compared state-of-art results are from NOC-REK [29]. Note that all the methods use the BERT-base[9] for a fair comparison. We compare our method with UpDown [5, 2], OSCAR [18], VinVL [35], VIVO [13], and NOC-REK [29], which hold the state-of-the-art result on the Nocaps benchmark. The training data for the baselines is the COCO dataset. Following prior settings, we also report the results after the model is optimized using

Table 1: Our evaluation results using SPICE and CIDEr on the Nocaps validation and test sets. We achieve the best scores for in-domain, near-domain, out-domain and Overall. Notably, the captions generated by our method are better than those by human in most cases. We note that our results on the test set are better than those by other methods which are publicly submitted to Nocaps leader-board^b. Higher score is better.

Method	Validation set								Test set							
	in-domain		near-domain		out-domain		Overall		in-domain		near-domain		out-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
UpDown [4]	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1	76.0	11.8	74.2	11.5	66.7	9.7	73.1	11.2
Oscar [18]	83.4	12.0	81.6	12.0	77.6	10.6	81.1	11.7	81.3	11.9	79.6	11.9	73.6	10.6	78.8	11.7
VIVO [13]	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	89.0	12.9	87.8	12.6	80.1	11.1	86.6	12.4
VinVL [35]	96.8	13.5	90.7	13.1	87.4	11.6	90.9	12.8	93.8	13.3	89.0	12.8	66.1	10.9	85.5	12.5
VinVL + VIVO [35, 13]	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	**78.0**	11.5	92.5	13.1
NOC-REK [29]	**104.7**	**14.8**	**100.2**	**14.1**	**100.7**	**13.0**	**100.9**	**14.0**	**100.0**	**14.1**	**95.7**	**13.6**	77.4	**11.6**	**93.0**	**13.4**
RCA-NOC**	95.0	13.6	91.3	13.1	93.2	12.1	92.2	13.0	87.7	13.3	88.1	12.9	77.9	11.4	86.3	12.7
RCA-NOC	**107.8**	**15.3**	**104.0**	**14.6**	**105.8**	**13.6**	**107.1**	**14.6**	**104.1**	**14.8**	**101.2**	**14.6**	**88.5**	**12.9**	**101.1**	**14.0**
Δ	3.1↑	0.5↑	4.2↑	0.5↑	5.1↑	0.6↑	6.2↑	0.6↑	4.1↑	0.7↑	5.5↑	1.0↑	11.1↑	1.3↑	8.1↑	0.6↑
Human [2]	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6

Table 2: Comparison of F1-scores (in %) on object classes of Open Images, evaluated on the Nocaps validation set. There are 504 classes in total. 80 of them are in-domain, which are common classes from COCO. The remaining 424 classes are the out-of-domain objects.

model	in-domain	out-of-domain	entire
VIVO	39.2	29.1	30.4
NOC-REK	45.3	30.5	32.8
Ours	58.6	39.2	43.1

Table 3: Evaluation results on the Held-Out COCO test set. The best results are highlighted in red.

Method	Avg. F1-score	SPICE	Meteor	CIDEr
NOC	48.8	–	21.4	–
NBT	48.5	15.7	22.8	77.0
DNOC	57.9	–	21.6	–
ZSC	29.8	14.2	21.9	–
ANOC	64.3	18.2	25.2	94.7
VinVL+VIVO	71.8	24.5	30.6	132.8
NOC-REK	**76.3**	**26.9**	**32.8**	**138.4**
RCA-NOC	**79.5**	**27.0**	**35.9**	**142.8**

SCST [24]. Our generated captions also adopt Constrained Beam Search (CBS) following [3].

In Table 1, we present our results of SPICE and CIDEr scores on the Nocaps validation and test sets. We also show the results of our different training stages in the middle of the table, where RCA-NOC** denotes our first training stage result and RCA-NOC is our final result after CIDEr optimization. Note that we do not compare with the data-driven methods[17, 30, 34] which use massive out-of-domain image-caption pairs in the training and fail to follow the rule of Nocaps. By leveraging relative semantic relevance in contrastive learning, our method has achieved a significant improvement compared to all prior works. It is worth noting that our first stage training could generate a comparable result with [13, 35, 18]. After the CIDEr optimization process, our method could outperform the recent state-of-art method NOC-REK [29] with a large margin, which is 8.1 and 0.6 (CIDEr and SPICE) better than [29] on the Nocaps Test set. This suggests that our model is more capable of generating captions with novel objects.

To quantitatively evaluate how well the model can describe novel objects, we also calculate the F1-score between our generated and the ground-truth tags on the validation set. Table 2 shows the comparison with VIVO and NOC-REK on the Nocaps validation set. We see that RCA-NOC improves NOC-REK and VIVO on F1-score substantially, especially for out-of-domain objects. This again verifies the effectiveness of RCA-NOC's discriminative ability to describe and distinguish novel objects.

Results on the Held-Out COCO dataset. To further prove the generalization ability of our method, we also conduct experiments on the Held-Out COCO dataset. As we can see from Table 3, our method consistently beats the baseline VinVL+VIVO with a large margin. The improvements are 7.7, 2.5, 5.3, and 10.0 with BERT-base on the F1, SPICE, METEOR, and CIDEr metrics. Additionally, our method is superior to all other state-of-art methods. In particular, compared with the recent state-of-art method NOC-REK [29], the recent state-of-art on the Held-Out COCO dataset, RCA-NOC achieves a 4.4 improvement on CIDEr (from 138.4 to 142.8) and 3.1 improvement on METEOR (from 32.8 to 35.9). The 3.2 improvement on the F1 score also shows that RCA-NOC performs better in describing novel objects.

4.3. Ablation study.

In this subsection, we will discuss every component's contribution to our framework. If not mentioned by purpose, all methods are conducted on the Nocaps vali-

Table 4: Effectiveness of the source of visual semantics.

Method	in-domain		near-domain		out-of-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
N.A.	103.6	14.7	100.8	14.2	100.9	13.2	103.2	14.2
NOC-REK	103.9	14.9	101.6	14.3	102.0	13.4	103.4	14.3
VIVO	105.2	15.1	103.5	14.4	103.6	13.4	105.2	14.4
CLIP	107.8	15.3	104.0	14.6	105.8	13.6	107.1	14.6

Table 5: Effectiveness of the UASR. UAR: Uncertainty-Aware Re-weight, UAS: Uncertainty-Aware Selection, UAS^N: Uncertainty-Aware Selection for Negatives), UASR: Uncertainty-Aware Selection and Re-weight.

Method	in-domain		near-domain		out-of-domain		Overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
N.A.	103.6	14.7	100.8	14.2	100.9	13.2	102.4	14.2
UAS^N	105.2	14.8	102.1	14.3	102.4	13.3	104.2	14.3
UAS	105.9	15.0	104.5	14.3	103.9	13.4	105.3	14.5
UAR	104.3	14.8	103.3	14.4	104.7	13.3	105.6	14.3
UASR	107.8	15.3	104.0	14.6	105.8	13.6	107.1	14.6

dation set with BERT-base. (we cannot use the Nocaps test set because only 5 times submissions are allowed for the test set).

Effectiveness of visual semantics. We enhance the modality alignment by incorporating visual semantics from foundational models, with a specific instantiation using CLIP. To evaluate the generalization ability of our approach, we modify the model to extract visual semantics and conduct further experiments to compare it with NOC-REK [29] and VIVO for generating visual semantics. Our baseline model is trained using a contrastive approach that ranks ROI tags extracted from Faster R-CNN based on their softmax values, distinguishing between relevant and irrelevant tags.

Our findings indicate that using NOC-REK to generate augmented tags provides only marginal improvement to our framework, resulting in a 0.8 and 0.1 (CIDEr and SPICE) increase compared to the baseline. We hypothesize that this limited improvement may be due to the restricted linguistic knowledge contained in BERT, which may not encompass a wide range of novel categories and lack higher-level and global concepts required to describe complex scenes. In contrast, utilizing a larger pool of noisy visual semantics (CLIP) extracted from captions leads to better results, with a 4.9 and 0.4 (CIDEr and SPICE) increase compared to the baseline. Using a smaller and cleaner pool of visual semantics extracted from the classes of various object detection datasets (VIVO) only results in a 2.0 and 0.2 (CIDEr and SPICE) increase compared to the baseline. This is likely due to the fact that the CLIP is pre-trained on vast amounts of training data, enabling it to extract a more diverse, extensive, and semantically meaningful set of high-level visual concepts compared to NOC-REK.

Effectiveness of the UASR. We investigate the performance of our Uncertainty-Aware Selection and Reweighting (UASR) on the Nocaps validation set in this part. Table 5 shows that Uncertainty-Aware Reweighting (UAR) could effectively enhance the reliable contrastive samples and bring 3.2 and 0.1 boosts on CIDEr and SPICE. In addition, the false positives/negatives severely harm our training process. By further incorporating Uncertainty-Aware Selection for Negatives (UAS^N) into our method and filtering false negatives, our model's performance is boosted to 104.2 and 14.3 on CIDEr and SPICE. When filtering false positives/negatives simultaneously, we could reach 105.3 and 14.5 on CIDEr and SPICE with Uncertainty-Aware Selection (UAS). Finally, our method improves the performance by 4.7 and 0.4 (CIDEr and SPICE) by combining UAR and UAS, which illustrates the proposed UASR is a more powerful tool for tackling noisy contrastive samples in NOC.

Table 6: Evaluation on COCO test set of Karpathy split [15]. All results are based on single model with cross-entropy optimization.

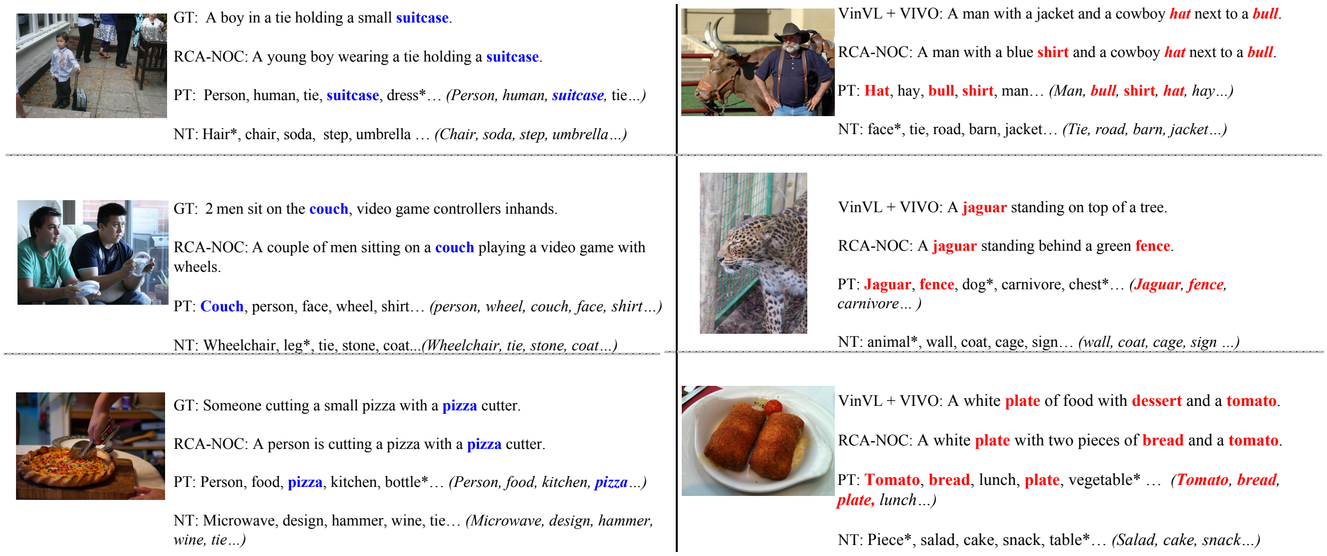
pre-training	BLEU4	Meteor	CIDEr	SPICE
N.A	33.8	28.1	118.3	21.2
OSCAR + VIVO	34.9	28.4	119.8	21.7
RCA-NOC	**37.4**	**29.6**	**128.4**	**23.1**

Table 7: Ablation study on the effectiveness of different components, including (i) CA: Cross-modality Alignment, our prototype contrastive loss for region-tag pairs, (ii) IA: Inner-modality Alignment, (iii) CLIP: Without CLIP refers to extracting visual semantics with object detection, and (iv) UASR: Uncertainty-Aware Selection and Re-weighting.

CA	IA	UASR	CLIP	CIDER	SPICE
				90.9	12.8
✓				96.5	13.5
✓	✓			98.7	13.8
✓	✓	✓		103.2	14.2
✓	✓	✓	✓	107.1	14.6

4.4. Qualitative Results

In Figure 3, we display some qualitative results on Held-Out COCO and Nocaps, and all the approaches are based on BERT-base. On Held-Out COCO, the result compared with GT shows that our proposed method could effectively generate accurate and precise captions with novel objects by leveraging relative semantic relevance into training. On Nocaps, VinVL+VIVO [35, 13] sometimes cannot include the desired novel objects in their captions (first and third examples) or generate wrong captions (second example). Our RCA-NOC, on the other hand, successfully generates correct and coherent captions via differentiating confusing classes and aligning object detection with other modalities. For better illustration, we show top-5 Positive and Negative Tag results extracted from CLIP.



GT: A boy in a tie holding a small **suitcase**.

RCA-NOC: A young boy wearing a tie holding a **suitcase**.

PT: Person, human, tie, **suitcase**, dress*… *(Person, human, **suitcase**,* tie*...)*

NT: Hair*, chair, soda, step, umbrella … *(Chair, soda, step, umbrella…)*

GT: 2 men sit on the **couch**, video game controllers inhands.

RCA-NOC: A couple of men sitting on a **couch** playing a video game with wheels.

PT: **Couch**, person, face, wheel, shirt… *(person, wheel, couch, face, shirt…)*

NT: Wheelchair, leg*, tie, stone, coat...*(Wheelchair, tie, stone, coat…)*

GT: Someone cutting a small pizza with a **pizza** cutter.

RCA-NOC: A person is cutting a pizza with a **pizza** cutter.

PT: Person, food, **pizza**, kitchen, bottle*… *(Person, food, kitchen, **pizza**…)*

NT: Microwave, design, hammer, wine, tie… *(Microwave, design, hammer, wine, tie…)*

VinVL + VIVO: A man with a jacket and a cowboy ***hat*** next to a ***bull***.

RCA-NOC: A man with a blue **shirt** and a cowboy ***hat*** next to a ***bull***.

PT: **Hat**, hay, **bull**, **shirt**, man… *(Man, **bull**, **shirt**, **hat**, hay...)*

NT: face*, tie, road, barn, jacket… *(Tie, road, barn, jacket…)*

VinVL + VIVO: A **jaguar** standing on top of a tree.

RCA-NOC: A **jaguar** standing behind a green **fence**.

PT: **Jaguar**, **fence**, dog*, carnivore, chest*… *(**Jaguar, fence,** carnivore…)*

NT: animal*, wall, coat, cage, sign… *(wall, coat, cage, sign ...)*

VinVL + VIVO: A white **plate** of food with **dessert** and a **tomato**.

RCA-NOC: A white **plate** with two pieces of **bread** and a **tomato**.

PT: **Tomato**, **bread**, lunch, **plate**, vegetable* … *(**Tomato, bread, plate,** lunch…)*

NT: Piece*, salad, cake, snack, table*… *(Salad, cake, snack…)*

Figure 3: Examples of generated captions and contrastive tags by compared methods on Held-Out COCO (left) and Nocaps (right). We show the ground-truth captions (GT) on Held-Out COCO for reference. PT/NT denotes the positive/negative tags before and after Uncertainty-Aware Selection and Reweighting (UASR). **Blue**/**Red** text indicates novel objects in Held-Out COCO/Nocaps, and $*$ indicates the false positives/negatives.

4.5. General Image Captioning

Improving modality alignment is a shared objective of general image captioning tasks. As demonstrated in Table 6, our proposed method, RCL-NOC, improves the model's performance across all metrics assessed on the COCO test set, particularly in the CIDEr score. However, we have observed that the improvement on the COCO benchmark is not as substantial as that on the nocaps benchmark. We hypothesize that this discrepancy is due to the COCO dataset having a limited number of visual concepts, thus reducing the benefits of learning visual semantics. Additionally, our method's use of tags is general and does not rely on fully annotated data. It's possible to utilize potentially unlimited amounts of images and different sources of tags, including those from Faster R-CNN, image tagging, or keywords extracted from captions. These possibilities will be explored in our future work..

4.6. Data and Compute Efficiency

The way the visual semantics are extracted (using CLIP) is not central to our work, as for the gain coming from 400M pairs of CLIP. This is supported in section 4.3. Table 7 shows that other components contribute significantly. Notably, the gain achieved by our Contrastive Alignment (CA) component alone can surpass that of CLIP by a large margin. Importantly, our method achieves this without increasing the number of parameters during training and inference, with the same parameter count as NOC-REK and VINVL (110 M). The training time of NOC-REK is 49 hours, for NOC-REK the training time is much lower (23h). This is with the paper setup (8 GPUs A100). Compared with NOC-REK, which adopts test time augmentation during inference, our inference time is much lower (1 minute) while NOC-REK is 2 minutes.

5. Conclusion

In this paper, we present RCA-NOC, which achieves visual-semantic alignment via relative contrastive learning. Specifically, for each image, we first extract augmented tags obtained from foundation models such as CLIP. Then we utilize the rank of each augmented object tag in a list as a relative relevance label to contrast each top ranked tag with a set of lower ranked tags. We empirically find that such a learning objective is effective and easy to optimize by encouraging the top ranked tags to be more compatible with their image and text context than lower ranked tags, hence improving the discriminative ability of the learned multimodality representation. We prove the effectiveness of our paradigm in novel object caption, with the spotlight on Nocaps and Held-Out COCO benchmark.

The effectiveness of object tags has already been proved in recent works. However, there are few studies to further explore why such tags could perform well and the problem of object tags' misalignment is often ignored. So how to further exploit use tags explicitly while having a deep understanding of their roles is meaningful in the future.

References

[1] https://huggingface.co/transformers/modeldoc/bert.html. 5

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 5, 6

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017. 6

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 6

[5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 5

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[7] Xianyu Chen, Ming Jiang, and Qi Zhao. Leveraging human attention in novel object captioning. In *IJCAI*, 2021. 1, 2

[8] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Image captioning with unseen objects. In *BMVC*, 2019. 1, 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 5

[10] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 1, 2, 3, 4

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[12] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 2, 5

[13] Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. Vivo: Visual vocabulary pre-training for novel object captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1575–1583, 2021. 1, 2, 3, 5, 6, 7

[14] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 7

[16] Jiacheng Li, Chang Chen, and Zhiwei Xiong. Contextual outpainting with object-level contrastive learning supplementary material. 2

[17] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 2, 3, 6

[18] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2, 3, 5, 6

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[20] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018. 1, 2

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[22] 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, 2021. 2, 3

[23] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[24] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 5, 6

[25] Mikihiro Tanaka and Tatsuya Harada. Captioning images with novel objects via online vocabulary expansion. In *ECCV*, 2020. 1

[26] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. Rich image captioning in the wild. In *CVPR Workshop on DeepVision*, 2016. 1

[27] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, 2017. 1, 2

[28] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1

[29] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge, 2022. 1, 2, 5, 6, 7

[30] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 6

[31] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[32] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021. 2

[33] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1029–1037, 2018. 1, 2

[34] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 6

[35] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 1, 2, 5, 6, 7