Paraphrasing Is All You Need for Novel Object Captioning

Cheng-Fu Yang[1] **Yao-Hung Hubert Tsai**[2] **Wan-Cyuan Fan**[3]

Ruslan Salakhutdinov[2] **Louis-Philippe Morency**[2] **Yu-Chiang Frank Wang**[3,4]

[1]UCLA [2]Carnegie Mellon University [3]National Taiwan University [4]NVIDIA

cfyang@cs.ucla.edu

Abstract

Novel object captioning (NOC) aims to describe images containing objects without observing their ground truth captions during training. Due to the absence of caption annotation, captioning models cannot be directly optimized via sequence-to-sequence training or CIDEr optimization. As a result, we present *Paraphrasing-to-Captioning (P2C)*, a two-stage learning framework for NOC, which would heuristically optimize the output captions via paraphrasing. With P2C, the captioning model first learns paraphrasing from a language model pre-trained on text-only corpus, allowing expansion of the word bank for improving linguistic fluency. To further enforce the output caption sufficiently describing the visual content of the input image, we perform self-paraphrasing for the captioning model with fidelity and adequacy objectives introduced. Since no ground truth captions are available for novel object images during training, our P2C leverages cross-modality (image-text) association modules to ensure the above caption characteristics can be properly preserved. In the experiments, we not only show that our P2C achieves state-of-the-art performances on nocaps and COCO Caption datasets, we also verify the effectiveness and flexibility of our learning framework by replacing language and cross-modality association models for NOC. Implementation details and code are available in the supplementary materials.

1 Introduction

Novel Object Captioning (NOC) [1] requires one to accurately describe images containing novel objects unseen during training captions. Despite impressive benchmark performance on COCO Captions [2] and Flickr30K [3], existing image captioning models [4, 5, 6, 7, 8] or unsupervised image captioning works [9, 10] cannot generalize well. This is because existing unsupervised captioning models typically assume that training image and caption data share the same visual content (i.e., objects) of interest, which might not be held in practice [9, 10].

Without observing captions describing novel objects during training, a number of NOC works choose to rely on object detection results for filling in the generated slotted sentences [11, 12]. Since the object words and the template sentences are generated separately, their relationships might not be well described. Therefore, Hu et al. [13] propose to relate images and the produced captions by aligning the region features of an object and its associated word embedding, aiming at improved description of novel objects in scenes which are similar to the training ones. However, if a word is not presented in the training corpus, it might not be properly presented in the predicted caption. That is, the word embedding related to novel objects such as verbs and adjectives might not be fully exploited

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

during inference, resulting in unsatisfactory *linguistic fluency* of output captions. On the other hand, despite the visual feature of a novel object is aligned with textual feature, existing methods generally are not designed to sufficiently caption images containing such objects of interest. In other words, the *fidelity* and *adequacy* of the output captions cannot be preserved.

In this paper, we uniquely approach the task of NOC by introducing and learning paraphrasing capabilities into state-of-the-art captioning models. More specifically, we advance pre-trained language models to expand the word bank of a captioning model for NOC, followed by enforcing the self-paraphrasing ability of this NOC model. The goal of the former stage is to preserve the linguistic fluency for NOC models, while that for the latter stage is deployed to exhibit improved fidelity and adequacy of the learned model. Since no ground truth captions of novel object images are available during training, we apply cross-modality association model with objectives/critics particularly designed for NOC.

It is worth noting that, for the evaluation part of this work, we not only show that our method achieves state-of-the-art performances on the nocaps and COCO Caption datasets, we further assess the metrics reflecting the fluency, fidelity, and adequacy of output captions. In addition, via ablation studies, we further confirm the practicality and flexibility of our learning framework, which does not limit to particular language or cross-modality association modules for paraphrasing and image-text alignment.

2 Related Work

Image captioning. Recent progress of image captioning focuses on different model architectures and learning methods. Huang et al. [4], Wang et al. [5], Guo et al. [6], Cornia et al. [7] design different attention mechanisms for image captioning. Rennie et al. [14], Li et al. [15], Yang et al. [16] adopt reinforcement learning to improve the performance. On the other hand, some researchers consider more challenging settings, such as partially supervised [17, 18] or unpaired image captioning [9, 10]. However, these methods are restricted to the assumption that the unpaired images and captions share the same set of object class, and the number of object class is limited as well, which make them inapplicable to our task.

Novel object captioning. Previously, novel object captioning approaches [19, 20, 21] were only tested on a restrictive dataset with only eight novel object classes held out from the COCO dataset. Their extensions to large-scale image data with various novel objects are not sufficiently studied. Recent studies mainly rely on object detection results to improve the performance on novel object captioning. Lu et al. [11], Wu et al. [12] generate slotted caption templates, which are later filled in with visual concepts identified by object detectors. Yao et al. [22] exploit a copying mechanism to assemble words corresponding to object detector predictions to generate captions. Similarly, Constrained Beam Search (CBS) [19] is an architecture-agnostic decoding algorithm that can be exploited during inference to enforce the inclusion of novel object classes in the captions. Instead of explicitly using detection results, Hu et al. [13] and Vo et al. [23] learn the relationship between image and text by aligning object detection tags with their corresponding image region features. Recently, Wang et al. [24] indicate that a desirable caption should comprise properties of fluency, fidelity, and adequacy. Nevertheless, most existing NOC approaches are not designed to handle language expression and cross-modal association with the above properties preserved.

Large-scale Vision and Language Pre-training (VLP). Recently, researchers discover that scaling up the sizes of both captioning model and training dataset would be effective to improve the performance on vision and language tasks. Dual-encoder frameworks such as CLIP [25] and ALIGN [26] scale-up contrastive pre-training [27] using 400M and 1.8B image-text pairs for cross-modal alignment. On the other hand, Transformer-based models like [28, 29] not only scale up the training corpus to 5.65M image and caption pairs, but also increase the transformer layers from 12 to 24. More recently, SimVLM [30] and LEMON [31] further explore the *huge* version of Transformer with a total of 32 layers, and scale up the pre-training corpus with 1.8B and 200M image-text pairs, respectively. Despite the dataset scale as an important factor in image captioning, we will demonstrate that our model still performs favorably against current large-scale methods under the same model size (i.e., Transformer *base* version, 12 layers) and with a smaller training caption corpus.

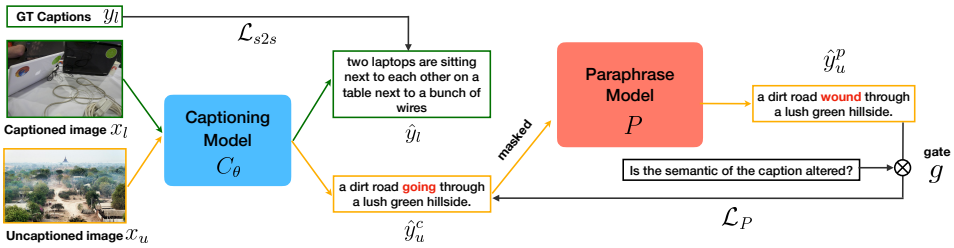

Figure 1: Learning to caption novel objects with linguistic fluency. For caption-labeled image x_l, we impose the sequence-to-sequence objective $\mathcal{L}_{s2s}$ for training. For uncaptioned image x_u, we exploit P to paraphrase the generated caption $\hat{y}_u^c$, producing the refined caption $\hat{y}_u^p$, followed by a semantic-preserving gate g to verify whether the paraphrased caption has altered the visual content.

3 Method

We first determine the notations and settings for the sake of completeness. Given a small set of images X_l with the corresponding captions Y_l, as well as a large set of uncaptioned images X_u containing novel objects, our goal is to generate the captions $\hat{Y}_u$ for X_u via a captioning model C_θ (θ describes the parameters for the captioning model C). To achieve this, we propose *Paraphrasing-to-Captioning (P2C)*, which allows C to perform paraphrasing to preserve linguistic fluency, and to self-paraphrase for boosting the associated fidelity and adequacy. The former is regularized by a pre-trained language model P, while an image-text cross-modality model A is utilized for guiding the unsupervised learning process. We note that the contribution of our proposed P2C lies in how to leverage the paraphrasing mechanism, together with the linguistic and visual information from pre-trained models P and A, for performing NOC with sufficient caption fluency, fidelity and adequacy.

3.1 Describing Novel Objects with Linguistic Fluency

By observing image-caption pairs (x_l, y_l), the captioning model C_θ in Fig. 1 would learn the visual grounding (i.e., localization of known objects and referring their expressions) as well the linguistics information. The training objective of such labelled data is a conventional sequence-to-sequence loss $\mathcal{L}_{s2s}$, calculated on ground-truth caption y_l. That is, we have $\mathcal{L}_{s2s} = \text{CrossEntropy}(\hat{y}_l, y_l)$.

However, using the image-caption pairs (x_l, y_l) alone is not sufficient to produce fluent captions for uncaptioned images X_u. We observe that some commonly-used wording of the associated novel objects, are not presented in the training corpus. Therefore, we leverage a language model P, to learn its linguistic knowledge via *paraphrasing*. That is, for uncaptioned images X_u containing novel objects, we first generate a caption describing X_u. Then, we leverage pre-trained language models as the paraphrase model P, which replace the generated captions with the most probable wording of a given novel object context. A semantic-preserving gate g is deployed to validate the paraphrased caption. As detailed later, this gate g is realized by a cross-modal association model A, which assesses the relationship between the paraphrased caption and the input image, ensuring semantics of the caption is not modified by P.

Learning to paraphrase via language model P. We now present the detailed paraphrasing processing for learning C_θ. As illustrated in Fig. 1, given an uncaptioned image x_u, the captioning model C_θ generates a caption $\hat{y}_u^c = C_\theta(x_u)$, $\hat{y}_u^c = \{w_1^c, w_2^c, ..., w_T^c\}$, where w_i^c denotes the ith word, and T is the caption length. The superscript c represents it is generated by our captioning model. We then obtain a masked caption $\hat{y}_u^M = \{w_1^c, w_2^c, ..., w_m^M, ..., w_T^c\}$ with m indicating the mask index, by randomly masking out the words in the caption. We note that, we do *not* mask nouns/objects in the above process, since they are related to objects grounded in the (novel) visual content and thus are not explicitly associated with caption fluency. As a result, P takes the masked sequence $\hat{y}_u^M$ as input and predicts the masked word w_m^p with the highest probability conditioned on the context of the entire sentence, producing the paraphrased caption $\hat{y}_u^p = P(\hat{y}_u^M)$, $\hat{y}_u^p = \{w_1^c, w_2^c, ..., w_m^p, ..., w_T^c\}$.

Semantic-preserving gate g. In the above paraphrasing process, however, not every word substitution from P is guaranteed to be semantically correct. We thus require a proper image-text model (i.e., association model) A to validate the paraphrasing output. That is, if the paraphrased caption comprises a more accurate and associated word that human generally uses to describe the scene, then a higher

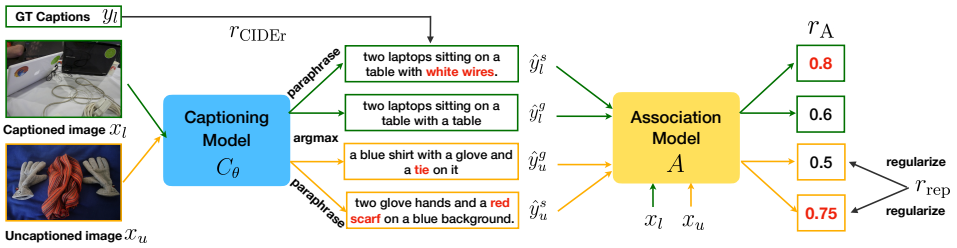

Figure 2: Improving caption fidelity and adequacy via self-paraphrasing. For caption-labeled image x_l, we perform CIDEr optimization. The sampled caption $\hat{y}_d^s$ will be rewarded by A if it has a higher cross-modal association than the greedy-decoded baseline $\hat{y}_d^g$. The superscript d indicates the source of the image. Additionally, we regularize our model with r_{rep} to avoid repetitive captions.

score would be obtained (than that of the original caption). Thus, we propose the objective function $\mathcal{L}_P$ to calculate the loss for the replaced words, gated by the comparison of the association scores of the corresponding captions. More precisely, $\mathcal{L}_P$ is derived as:

$$\mathcal{L}_P = -g * t \log(s), \quad g = \max(0, \tanh(A(x_u, y_u^p) - A(x_u, y_u^c) - \alpha)) \tag{1}$$

where g is the gating function preventing C_θ from learning from low-quality paraphrased captions, t is the one-hot representation of the paraphrased word, and s is the word distribution predicted by the captioning model at the masked timestep m, i.e., we have $w_m^c = \arg\max(s)$. We have $A(x, y)$ in g to calculate the association between an image x and its caption y, indicating how well the captions match the images, and α serves as the margin of the association scores for distinguishing between the two captions.

3.2 Improving Caption Fidelity and Adequacy via Self-Paraphrasing

Recall that, as discussed in Sec. 1, caption fidelity verifies whether the details of visual content are presented in the generated caption, while adequacy assesses whether the visual content is properly described been expressed in it. While the previous paraphrasing stage increases the linguistic fluency for captions describing images containing novel objects, the novel object word itself still can be incorrectly predicted or missing in the output captions. This is because that, the caption corpus Y_l does not contain any novel objects, and thus the novel object words have low probability to be captioned.

To tackle the above issues, one possible solution is to *self-paraphrasing* for C_θ with captioning evaluation metric as a critic to reward the paraphrased captions. For labeled images X_l, we can exploit the CIDEr score [14], encouraging the generated caption to be consistent with that of the human annotated ones in the word level. However, while CIDEr can be easily computed for captioning labeled images X_l, it cannot be explicitly calculated for captioning uncaptioned images X_u due to the absence of ground-truth captions. To address this problem, we discover that cross-modal association is an ideal measurement to reflect the fidelity and adequacy for output captions (see remarks in Appendix A). Specifically, we utilize an association model A to compute the association score between images and the generated captions, which serves as a *label-free* critic in our learning framework. Then, when a paraphrased caption is rewarded by a higher association score, our P2C would increase the probability of using that word in the sentence to describe the image. This learning strategy allows producing captions that precisely describe the objects with plentiful visual details.

Rewarding the generated captions. As a potential challenge in NOC, we observe that captioning models would achieve improved association by simply repeating the same object that occurs in the image, which undermines the linguistic fluency of the captions. For example, the image caption *"a group of cans of soda and other items on a table"* can be replaced by *"a pile of **cans** and bottles of soda on a counter with **cans** of **cans**"* with a higher association score.

To tackle this problem, we choose to impose a repetition penalty to avoid such trivial solutions. For image-caption pairs (x_l, y_l), we directly calculate the CIDEr reward for the predicted caption $\hat{y}_l$ (i.e., $r_{\text{CIDEr}} = \text{CIDEr}(\hat{y}_l, y_l)$). To enforce the generated captions for $(X, Y) = (X_l, Y_l) \cup (X_u, Y_u)$ with sufficient fidelity and adequacy, we exploit A to compute the association reward between X and $\hat{Y}$ (i.e., $r_{\text{A}} = A(x, \hat{y})$). As for the repetition penalty to preserve linguistic fluency of the generated captions $\hat{y}_u = \{w1, w2, ..., w_T\}$ for x_u, we formulate it as a linear assignment problem, where every

word is assigned to the most similar one in the same sentence except for itself. Then, we calculate the similarity between such pairs for each sentence. Intuitively, repetitive captions would have high similarity scores (with repeating words assigned to the exact same word)s. Thus, we define the assignment $\hat{\alpha}$ as the one that maximizes the average pairwise similarity of a sentence, i.e.,

$$\hat{\alpha} = \arg\max_{\alpha} \frac{1}{T} \sum_{i=1}^{T} C(w_i, w_{\alpha(i)}), \tag{2}$$

where $\alpha(i)$ is the index of the word assigned to the i-th word in the caption, and $C(w_i, w_j)$ is the cosine similarity between the GloVe [32] word representation of two words. Since a desirable captioning model would encourage captions with low repetition (i.e., low average pairwise similarity), the reward for repetition penalty is defined as follows:

$$r_{\text{rep}} = 1 - \frac{1}{T} \sum_{i=1}^{T} C(w_i, w_{\hat{\alpha}(i)}), \tag{3}$$

Note that we do not calculate repetition penalty for the labeled data $\hat{Y}_l$ since they are regularized by the aforementioned CIDEr rewards.

With the above discussions, the total reward for caption-labeled data would be $r(\hat{y}_l) = r_{\text{CIDEr}}(\hat{y}_l, y_l) + r_{\text{A}}(x_l, \hat{y}_l)$, and the total reward for uncaptioned data would be $r(\hat{y}_u) = r_{\text{A}}(x_u, \hat{y}_u) + r_{\text{rep}}(\hat{y}_u)$.

Back-propagation via reinforce algorithm. Unfortunately, computation of the aforementioned rewards is non-differentiable. Thus, we adopt reinforce algorithm [33] to optimize our P2C learning. As shown in Fig. 2, for an image x we use greedy decoding to obtain the baseline result $\hat{y}^g$, and the paraphrased captions $\hat{y}^s$ are derived from randomly sampling from the word distribution. If the sampled captions possess higher linguistic fluency or cross-modal association than the baseline caption, they will be encouraged by positive rewards and vice versa. We follow Rennie et al. [14], Liu et al. [34, 17] and define the objective function as follows:

$$\begin{aligned} \nabla_\theta \mathcal{L}_{RL}(\theta) &\approx -(r(\hat{y}_d^s) - r(\hat{y}_d^g)) \nabla_\theta \log p_\theta(\hat{y}_d^s), \\ r(\hat{y}_d) &= \begin{cases} r_{\text{CIDEr}}(\hat{y}_d, y_d) + r_{\text{A}}(x_d, \hat{y}_d) & \text{if } x_d \in X_l \\ r_{\text{A}}(x_d, \hat{y}_d) + r_{\text{rep}}(\hat{y}_d) & \text{if } x_d \in X_u \end{cases}, \end{aligned} \tag{4}$$

where d indicates the source of the image, θ being the parameters of captioning model, and $p_\theta(\hat{y}^s)$ represents the predicted word logits for the generated captions. With the objective functions defined in equations (1), and (4), the NOC model C_θ can be trained accordingly.

4 Experiments

Implementation details. Following Hu et al. [13], Li et al. [28], Zhang et al. [29], we consider a BERT-base [35] architecture for our captioning model C_θ. To demonstrate the flexibility of our learning framework, we apply two different versions of BERT, *base* and *large*, which are pre-trained on large-scale *text*-only corpus to model human language as our paraphrase model P. For the association model A, we have three different cross-modal association models, VIFIDEL [36], SR-PL [17], and CLIP [25] with its version being ViT/B-32. VIFIDEL associates image-caption data using word embedding of particular object labels, and SR-PL utilizes the triplet loss to learn the association of image captions, while CLIP is optimized via the contrastive pre-training. For our model reported in the following subsections, we use BERT large for P, and CLIP for A. Due to page limits, hyperparameters and other training details can be found in Appendix B.

Datasets. The training data for the **nocaps** benchmark comprises the Open Images V4 [37] object detection training set (1.7M images annotated with bounding boxes for 600 object classes), plus the image-caption pairs from the COCO Captions 2017 [2] training set (0.5M image-caption pairs containing 80 object classes). No additional image-caption pairs are provided for training. We evaluate our model on the validation and test set of nocaps, which comprises 4500 and 10600 images from the Open Images validation and test sets, respectively. To compare with current large-scale methods and investigate our model performance when scaling up the training dataset, we jointly trained our model on the Conceptual Caption dataset [38] and compare our method to other methods that access additional image-text pairs during their training on the nocaps-XD [1] benchmark (XD

Table 1: Quantitative results on nocaps. Note that C and S denote CIDEr and SPICE, respectively. We highlight the highest score in blue, while the second best scores are marked in bold.

Method	in-domain		near-domain		out-domain		overall		in-domain		near-domain		out-domain		overall	
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
	Validation Set								Test Set							
UpDown	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2	76.0	11.8	74.2	11.5	66.7	9.7	73.1	11.2
Oscar$_B$	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
Oscar$_L$	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4	84.8	12.1	82.1	11.5	73.8	9.7	80.9	11.3
Oscar$_B$ +VIVO	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	**96.8**	13.5	90.7	13.1	87.4	11.6	90.9	12.8	93.8	13.3	89.0	12.7	66.1	10.9	85.5	12.5
VinVL +VIVO	94.8	13.3	**91.4**	13.0	88.7	11.6	**91.4**	12.7	**94.5**	13.1	**90.9**	12.9	77.9	11.3	**88.3**	12.6
Human	84.8	**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**
Ours	**101.4**	**15.1**	**96.8**	**14.5**	**95.4**	**12.9**	**97.2**	**14.2**	**101.7**	**15.0**	**95.7**	**14.4**	**82.5**	**12.2**	**93.5**	**14.1**

stands for extra data). In addition, to demonstrate that our method also enhances model performance on the general image captioning, we evaluate our P2C on the COCO Captions dataset. Due to page limits, the comparison on COCO Captions can be found in Appendix C.2.

4.1 Evaluation metrics

CIDEr. Like the evaluation metrics [39, 40, 41] in NLP, Consensus-based Image Description Evaluation (CIDEr) calculates the similarity between the reference and generated caption by n-gram overlap in a rule-based manner. To capture human consensus in image captioning, it introduces the tf-idf weight to reduce the matching weight of the n-grams that are common in all image captions.

SPICE. Semantic Propositional Image Caption Evaluation (SPICE) [42] matches the semantics between sentences, such as objects, relations, and attributes of objects. Specifically, it converts sentences into semantic scene graphs, which allows evaluation to break grammatical constraints and focuses on propositional semantic content. It reflects the accuracy of the visual content and considers less about linguistic properties.

Fluency. To quantitatively evaluate fluency, we remove the effect of the visual information and focus on the quality of linguistic properties in the conventional caption evaluation metrics. Specifically, we we disregard the n-grams containing the particular object word of interest during the computation BLEU@4 [39] and CIDEr scores. Take the sequence "a b c d" for example, when 'b' is the object word, only the unigram "a, c, d" and the bigram "cd" would be taken into account, n-grams such as "ab" or "abc" would be excluded from computation. Note that the fluency experiment is conducted on a subset of the nocaps validation set, which contains 1000 images whose caption annotations are available on the official website of the nocaps dataset.

Fidelity & Adequacy. Fidelity and adequacy evaluate how well the captions are associated with images. As defined in Sec. 1, fidelity evaluates whether the objects described by the caption are actually presented in the images, while adequacy assesses how many objects in the images are described by the captions. These two properties are analogous to the definition of precision and recall, respectively. Therefore, we extract the objects mentioned in the captions and the ground-truth objects in the images and calculate the precision (for fidelity), recall (for adequacy), and F1 (for overall association) scores. The experiment is performed on the validation set of nocaps.

Following Agrawal et al. [1], we split the dataset into three subsets for evaluation: *in-domain* images only contain the seen objects that have been described during training, *out-of-domain* images are the unseen/uncaptioned (i.e., novel) ones, while *near-domain* ones contain both seen and unseen objects.

4.2 Quantitative analysis

For performance comparisons, we choose UpDown [1] as baselines, as well as Oscar [28] and VinVL [29] which achieves SOTA results on the benchmark of nocaps. All methods are trained via the SCST optimization [14] except for the UpDown baseline, and Constrained Beam Search (CBS) is exploited during inference. In addition, VIVO [13] is a pre-training technique for captioning models, allowing them to recognize the novel objects. Since VinVL did not report the numbers with CBS exploited during inference (CBS is known to improve model performance on out-of-domain data), we reproduce VinVL following details stated in the original paper. For more details please refer to Appendix B. For a comprehensive comparison, we conduct experiments on the validation and test set of nocaps. In addition, we compare our method with VinVL and VIVO in fluency, fidelity,

Table 2: Quantitative results on the nocaps (XD) benchmark. Note that XD stands for extra data.

Method	Pre-training data	Validation set		Test set	
		CIDEr	SPICE	CIDEr	SPICE
Encoder-Decoder [43]	CC12M [43]	87.4	11.8	85.3	11.8
Encoder-Decoder	CC3M+CC12M	90.2	12.1	87.3	12.0
VinVL$_{base}$ [29]	5.65M Combined	95.5	13.5	-	-
SimVLM$_{base}$ [30]	1.8B	-	-	**94.8**	13.1
LEMON$_{base}$ [31]	CC3M [38]	91.6	13.0	-	-
LEMON$_{base}$	CC12M	100.4	13.8	-	-
LEMON$_{base}$	ALT200M [31]	**106.8**	14.1	-	-
Ours	COCO Caption 0.5M	97.2	**14.2**	93.5	**14.1**
Ours	CC3M	**104.1**	**14.6**	**102.4**	**14.7**

Table 3: Quantitative comparisons on caption fluency, fidelity and adequacy. Note that BLEU@4 (B@4) and CIDEr (C) are utilized for describing fluency, object precision (P) for fidelity, object recall (R) for adequacy and object F1 scores (F1) for overall cross-modal association.

Method	in-domain					near-domain					out-of-domain				
	B@4	C	P	R	F1	B@4	C	P	R	F1	B@4	C	P	R	F1
VinVL	21.6	74.8	**59.2**	40.8	48.3	19.6	73.3	22.8	32.6	26.8	17.9	59.6	48.4	25.6	33.5
VinVL+VIVO	20.6	71.6	56.0	42.2	48.1	19.8	75.4	28.5	36.3	32.0	17.4	59.6	49.0	27.3	35.1
Ours	**25.4**	**87.0**	58.2	**45.6**	**51.3**	**22.1**	**80.1**	**39.9**	**41.0**	**40.4**	**19.7**	**67.7**	**51.3**	**30.5**	**38.3**

and adequacy to demonstrate the improvement in terms of these properties. We also evaluate on the COCO Caption dataset and report in Appendix C.2.

The nocaps datasets. The results on nocaps are shown in Table 1. From this table, we see that our model performed favorably against baselines and SOTAs across different metrics. We see that our method substantially improves the CIDEr scores in every domain, which verifies our design to generate more fluent and natural captions. Also, it is worth noting that our model largely increased the performance in SPICE score for every data domain, which verifies that our method is able to generate captions with the improved image-language association.

The nocaps-XD datasets. To compare with large-scale pre-training methods and investigate our model performance when scaling up training dataset, we evaluate our method on the nocaps-XD benchmark and show the results in Table 2. Note that both VinVL, SimVLM, LEMON, and our method adopt BERT-based backbones for captioning. From the above table, we see that our method still performs favorably against SOTAs on the benchmark even if we have the fewest training caption samples (3M). Our model performance is only slightly below LEMON on the CIDEr score on validation set when it uses a larger training corpus (ALT200M), while we surpass LEMON by a significant margin when the same training corpus (CC3M) is used. Thus, the effectiveness of our method when we scale up the training data can be verified. For more detailed discussion and the ablation study on this model, please refer to Appendix C.3.

Fluency, fidelity, and adequacy. As described in Sec. 4.1, we design additional experiments for evaluating fluency, fidelity, and adequacy and report the results in Table 3. For fluency, we remove all the objects and nouns in the captions since they relate less to the linguistics of the captions. We then calculate the BLEU@4 (B@4) and CIDEr (C) scores for the captions after removal. For fidelity and adequacy, they indicate that captions should accurately (high precision) describe sufficient (high recall) visual details. Therefore, we report the object precision and recall in this table, and object F1 scores represent the overall association between captions and images. One can see that our method surpasses previous methods by a visible margin on all tasks except for in-domain object precision, which further verifies our model improves novel object captioning on fluency, fidelity, and adequacy.

4.3 Ablation studies

Following the same evaluation procedures in Sec. 4.1, we discuss the contributions of the uses of paraphrasing model P and association model A in terms of linguistic and semantic level metrics, and present their results in Table 4 and 5. In addition, to verify the flexibility of our proposed P2C, we replace P and A with different implementations of language models and association models. Then, we evaluate their performance on the nocaps validation set and report the results in Table 6. Detailed ablation analysis of every objective can be further found in Appendix C.1.

Table 4: Analyses on paraphrasing model P, association model A, and repetition penalty for NOC using nocaps validation set. Note that P mainly benefits the linguistic fluency with improved CIDEr, and reward from A is desirable for preserving visual semantics with increased SPICE.

Method	in-domain		near-domain		out-domain		overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Ours	**102.8**	**14.8**	**97.9**	**14.4**	**86.3**	**12.5**	**96.3**	**14.1**
Ours w/o g	32.8	10.6	21.5	9.5	12.6	8.0	21.3	9.4
Ours w/o $\mathcal{L}_P$	99.1	14.4	94.7	14.1	84.5	12.4	93.3	13.8
Ours w/o r_{A}	101.1	13.8	94.1	13.4	80.5	11.9	92.3	13.1
Ours w/o r_{rep}	96.7	**14.8**	89.6	14.1	81.9	12.4	89.1	13.9

Table 5: Analyses on paraphrase model P and association model A for improving caption fluency, fidelity and adequacy. Note that P benefits fluency metrics of BLEU@4 and CIDEr, while A focusing on cross-modal association boosts metrics of object precision, recall, and F1 scores.

Method	in-domain					near-domain					out-of-domain				
	B@4	C	P	R	F1	B@4	C	P	R	F1	B@4	C	P	R	F1
Ours	**25.4**	**87.0**	58.2	**45.6**	**51.3**	**22.1**	**80.1**	39.9	**41.0**	**40.4**	**19.7**	**67.7**	51.3	**30.5**	**38.3**
Ours w/o $\mathcal{L}_P$	21.6	77.1	58.1	40.9	48	19.8	75.5	**41**	39.0	40.0	19.1	66.1	**51.6**	19.1	66.1
Ours w/o r_{A}	23.0	79.3	**58.8**	42.1	49.1	21.2	78.6	35.8	37.7	36.8	18.8	65.7	**51.6**	27.4	35.8

Paraphrase model P. As shown in Table 4, the captioning model without P would observe a performance drop in CIDEr for linguistic fluency, but such drops for the metric of SPICE (related to visual content) would be less significant. Similarly, as observed in Table **??**, removing P would result in the lowest BLEU and CIDEr scores. These results confirm our motivation and model design, since P is utilized to improve caption quality at the linguistics level.

To further verify the use and flexibility of paraphrase model P, we replace our paraphrase model P (the *large* version of BERT) with the *base* version of BERT. We found out that only a slight performance drop is produced. We conjecture that this is because the two versions of BERT are trained on the same text corpus, which means they share the same word bank. As a result, our model would distill similar linguistic knowledge when using either model to guide the training of our P2C.

Semantic-preserving gate g. From Table 4, we see that captioning model would be misled by the wrong guidance produced by P, if no validation from g to ensure that the semantics of the original captions is not modified by P. This results in a significant performance drop on nocaps.

Association model A. As shown in Table 4, the model without the reward calculated by A showed significant drops in captioning metrics of CIDEr and SPICE. However, the performance decrease in SPICE is expected, since A is particularly deployed for preserving visual content in captions. As for CIDEr, its decrease is mainly due to the deterioration of missing visual content in captions. This is also confirmed by Table 5, in which mainly the metrics reflecting fidelity and adequacy would observe significant drops for model trained without the association rewards.

To further verify the use and flexibility of the association model A, we replace it with different models that also produce association scores between images and captions. Specifically, we consider VIFIDEL [36], SR-PL [17], and CLIP [25]. As the results shown in Table 6, one can see that the use of CLIP as A would achieve the best performance. This is because that, VIFIDEL only associates image-caption data using word embedding of particular object labels, while CLIP assesses such cross-modal data in the instance level, i.e., taking the complete caption of an image into consideration. As for SR-PL, it utilizes the triplet loss to learn the association of images and captions, while CLIP is optimized via the contrastive pre-training, which fully exploits the negative samples in a mini-batch to learn a more compact representation space, allowing it to estimate the association more accurately.

Repetition penalty. Recall that, in Sec. 3.2, this penalty is to alleviate the association between images and captions with redundant visual information. As seen in Table 4, the model without this penalty observed a significant performance drop in CIDEr scores. We did not see such trends for SPICE. This is because that, repetitive words in captions mainly violate linguistic structures rather then semantic accuracy, and thus the performance related to linguistic fluency would be more sensitive to the deployment of this penalty.

Table 6: Uses of different paraphrase models P and association models A for NOC. Results on the nocaps validation set are presented. Note that Stages 1 and 2 are the paraphrasing schemes presented in Sec. 3.1 and Sec. 3.2, respectively.

Method	Paraphrase Model P	Association Model A	$\mathcal{L}_P$	r_{A}	r_{rep}	Overall	
						CIDEr	SPICE
Baseline	N/A	N/A	✗	✗	✗	81.2	12.3
Ours w/ Stage 1	BERT_{base}	CLIP [25]	✓	✗	✗	84.1	12.7
Ours w/ Stage 1	BERT_{large}	CLIP	✓	✗	✗	84.2	12.7
Ours w/ Stage 2	BERT_{large}	VIFIDEL [36]	✓	✓	✓	55.3	10.7
Ours w/ Stage 2	BERT_{large}	SR-PL [17]	✓	✓	✓	84.6	13.2
Ours w/ Stage 2	BERT_{large}	CLIP	✓	✓	✓	**96.3**	**14.1**

	Fluency	Fidelity	Adequacy
VinVL	a group of women **standing** on a field with a football.	a **blue blanket** and **a glove** on a blue shirt.	a man and a woman playing an accordion.
Ours	a group of girls in football uniforms **posing** for a picture.	**two glove hands** and a **red scarf** on a blue background.	a couple of people **sitting in a church** playing an accordion.

Figure 3: Example results and comparisons for image captions produced by VinVL and ours in terms of fluency, fidelity and adequacy. Note that both utilize VIVO for novel object detection.

4.4 Qualitative analysis

For qualitative comparisons, we first conduct human study and ask individuals to evaluate a caption from three perspectives: fluency, fidelity, and adequacy. Due to page limit, the experiment details and the corresponding results are presented in Appendix C.4. For qualitative analysis, we empirically show captions in Fig. 3, which are generated by our model and VinVL, with both pre-trained from VIVO for novel object detection. In this figure, wordings that are less accurate or incorrectly describe the associated visual content are marked in bold. And, our wording improvements are highlighted in red. From this figure, one can see that for fluency, our model generated vivid captions with more proper wordings. Take the image on the left for example, our model particularly described *"posing for a picture"* instead of *"standing on a field"*. As for fidelity, our model is designed to accurately capture the visual content in an image. Specifically, take the second image for example, we correctly described the number of gloves and the novel object *red scarf*, while VinVL failed to do so. As for adequacy, take the bottom-right image for example, our model was able to recover visual details in the image (i.e., *"people playing the accordion"* and *"sitting in a church"*). For more qualitative examples, please refer to Appendix C.5.

5 Conclusion

In this paper, we proposed Paraphrasing-to-Captioning (P2C) for novel object captioning, with particular goals of improving caption fluency, fidelity, and adequacy. In P2C, we advocate the learning of two paraphrasing capabilities for captioning models. The first is language-level paraphrasing, which expands the word bank of a captioning model under the guidance of pre-trained language model, and thus preserves the linguistic fluency during NOC. Secondly, we introduce the self-paraphrasing ability for the captioning model to sufficiently describe visual content of the input image, so that both caption fidelity and adequacy can be achieved. Due to the lack of ground truth captions, image-text association is uniquely exploited for guiding the training process. Empirically, we not only showed that our model achieved SOTA results on benchmark datasets, we also assessed the metrics associated with fluency, fidelity, and adequacy, confirming the effectiveness of our model. Via our ablation studies, we further verified the flexibility of our learning framework by replacing language and cross-modality association modules for paraphrasing and image-text alignment. Finally, since we

apply pre-trained language and association models in P2C, their joint optimization would be among our future research directions.

Acknowledgement This work is supported in part by the Ministry of Science and Technology of Taiwan under grant NSTC 110-2634-F-002-052. Y.-H.H.T., R.S., and L.-P.M. was supported in part by the NSF IIS1763562, NSF Awards #1750439 #1722822, National Institutes of Health, IARPA D17PC00340, ONR Grant N000141812861, and Facebook PhD Fellowship. We also thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

[1] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019.

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

[3] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.

[4] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019.

[5] Weixuan Wang, Zhihong Chen, and Haifeng Hu. Hierarchical attention network for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8957–8964, 2019.

[6] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020.

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

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

[9] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332, 2019.

[10] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019.

[11] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018.

[12] 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.

[13] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. *arXiv e-prints*, pages arXiv–2009, 2020.

[14] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

[15] Nannan Li, Zhenzhong Chen, and Shan Liu. Meta learning for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8626–8633, 2019.

[16] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 1–17. Springer, 2020.

[17] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 338–354, 2018.

[18] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. *arXiv preprint arXiv:1909.02201*, 2019.

[19] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016.

[20] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. *arXiv preprint arXiv:1806.06004*, 2018.

[21] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016.

[22] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6580–6588, 2017.

[23] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge. *arXiv preprint arXiv:2203.14499*, 2022.

[24] Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059, 2021.

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

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

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

[28] 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.

[29] 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.

[30] 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.

[31] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*, 2021.

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[33] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[34] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[36] Pranava Madhyastha, Josiah Wang, and Lucia Specia. Vifidel: Evaluating the visual fidelity of image descriptions. *arXiv preprint arXiv:1907.09340*, 2019.

[37] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, 2018.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[40] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[41] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[42] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[43] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[44] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2020.

[45] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In our supplementary materials, we demonstrate some failure cases of our model. We empirically observe that the error mainly comes from the wrong detection tags predicted by the pre-trained object detectors.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...final
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see the supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 4. All datasets we used are public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Please see our supplementary materials.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]

Algorithm 1: Learning to caption novel objects with linguistic fluency

Input: Captioning model $C_\theta(\cdot)$, Paraphrase model P, and Association model A.
Data: Captioned image x_l, the corresponding GT caption y_l, uncaptioned image x_u, and lr η_{it}.
Output: Trained Captioning model $C_\theta(\cdot)$.

Initialize $C_\theta(\cdot)$;
for *it from 1 to num_iters* **do**
  $\hat{y}_l \leftarrow C_\theta(x_l)$, $\hat{y}_u^c \leftarrow C_\theta(x_u)$;
  Produce $\hat{y}_u^m$ by randomly masking words in the sentence in $\hat{y}_u^c$ (except for nouns);
  $\hat{y}_u^p \leftarrow$ P($\hat{y}_u^m$)
  $\mathcal{L}_{s2s} \leftarrow$ CrossEntropy($\hat{y}_l, y_l$)
  if $A(x_u, \hat{y}_u^b) \leq A(x_u, \hat{y}_u^c)$ **then**
    $\mathcal{L}_P \leftarrow 0$
  else
    $\mathcal{L}_P \leftarrow$ CrossEntropy($\hat{y}_u^c, \hat{y}_u^p$)
  end
  $\mathcal{L} \leftarrow \mathcal{L}_{s2s} + \mathcal{L}_P$
  Update parameters: $\theta \leftarrow Adam(\theta, \eta_{it}, \nabla_\theta \mathcal{L})$
end

A Remarks on fluency, fidelity, and adequacy

We first discuss how fluency, fidelity, and adequacy can be fundamentally and technically related to language model and association model. For caption fluency, one would expect that the image caption to be linguistically natural and fluent. That is, the NOC model not only requires to capture the occurrence of novel-object vocabularies, the associated collocations such as verbs or modifiers are expected to be properly utilized. Thus, given the context containing novel-object vocabularies as $\tilde{\mathrm{y}}$ and the associated wordings as $\hat{\mathrm{w}}$, we define fluency as the probability of the NOC model which correctly predicts the collocation given the novel image context $p(\hat{\mathrm{w}}|\tilde{\mathrm{y}})$ (N as the number of collocations). In natural language processing, masked language models are widely applied to predict the masked word $\hat{\mathrm{w}}$ given the context $\tilde{\mathrm{y}}$. Thus, the objective of a masked language model would be maximizing the log-likelihood of the masked word $\log(p(\hat{\mathrm{w}}|\tilde{\mathrm{y}}))$ given the context $\tilde{\mathrm{y}}$, which is equivalent to our definition of fluency with the log function explicitly calculated. Therefore, this is the reason why we adopt language model to learn the co-occurrence of novel-object vocabularies and their associated collocations to improve the linguistic fluency.

We now relate fidelity and adequacy in image captions to cross-modal association. We start by defining the probability of an object appearing in the images as $p(x)$, and the probability of an object mentioned by the captions as $p(y)$. Relevant objects $p(x, y)$ are defined as objects that are both included in the image and described by the associated caption. Since fidelity assesses whether the visual content presented in the produced caption is correct, it can be defined as the fraction of relevant objects among objects in captions $\frac{p(x,y)}{p(y)} = p(x|y)$. On the other hand, adequacy evaluates whether sufficient visual details have been expressed by captions, and it can be defined as the fraction of relevant objects among objects in an image $\frac{p(x,y)}{p(x)} = p(y|x)$. Thus, we can calculate the point-wise mutual information pmi of an image x and its caption y as follows:

$$pmi(x, y) \equiv log\frac{p(x, y)}{p(x)p(y)} = log\frac{p(x|y)}{p(x)} = log\frac{p(y|x)}{p(y)}, \tag{5}$$

with mutual information as the expected value of point-wise mutual information. For the task of NOC, both $p(x)$ and $p(y)$ are fixed (i.e., determined by the dataset). Therefore, the above derivation implies that when the mutual information between an image and its captions increases, the resulting fidelity and adequacy would be jointly improved. So that's why we require CLIP to compute the association between images and captions. Since it is trained via the InfoNCE objective, which is a lower bound estimation of mutual information [27].

Algorithm 2: Learning novel object captions with fidelity and adequacy

Input: Captioning model $C_\theta(\cdot)$ and Association model A.
Data: Captioned image x_l, the corresponding GT caption y_l, uncaptioned image x_u, and lr η_{it}.
Output: Trained Captioning model $C_\theta(\cdot)$.

Initialize $C_\theta(\cdot)$;
for *it from 1 to num_iters* **do**
  $\hat{y}_l^s \leftarrow C_\theta(x_l),\ \hat{y}_u^s \leftarrow C_\theta(x_u)$ (by sampling);
  $\hat{y}_l^g \leftarrow C_\theta(x_l),\ \hat{y}_u^g \leftarrow C_\theta(x_u)$ (by greedy decoding);
  Calculate $r_{\text{rep}}(\hat{y}_u^s)$ and $r_{rep}(\hat{y}_u^g)$ by (3)
  $r(\hat{y}_l) \leftarrow r_{\text{CIDEr}}(\hat{y}_l, y_l)$ + $r_{\text{A}}(x_l, \hat{y}_l)$
  $r(\hat{y}_u) \leftarrow r_{\text{A}}(x_u, \hat{y}_u)$ + $r_{\text{rep}}(\hat{y}_u)$
  Calculate the gradient $\nabla_\theta \mathcal{L}_{RL}(\theta) \leftarrow -(r(\hat{y}_d^s) - r(\hat{y}_d^g))\nabla_\theta \log p_\theta(\hat{y}_d^s),\ d \in \{l, u\}$
  Update parameters: $\theta \leftarrow \mathit{Adam}(\theta,\ \eta_{it},\ \nabla_\theta \mathcal{L}_{RL})$
end

Table 7: Ablation studies on nocaps validation set.

Method	in-domain		near-domain		out-of-domain		overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Baseline (Only w/ L_{s2s})	89.07	13.29	83.29	12.61	68.77	10.59	81.17	12.32
+ L_P	92.46	13.40	85.79	12.92	73.21	11.40	84.20	12.69
+ r_{CIDEr}	101.19	13.84	95.38	13.44	83.24	12.06	93.75	13.23
+ r_{A}	96.73	**14.83**	89.64	14.12	81.87	12.38	89.08	13.88
+ r_{rep} (Ours)	**102.77**	**14.83**	**97.90**	**14.40**	**86.33**	**12.54**	**96.25**	**14.10**

B Implementation details

Following Hu et al. [13], Li et al. [28], Zhang et al. [29], we consider a BERT-base [35] architecture for our captioning model. Given an image, the captioning model jointly takes the image region features and the predicted detection tags to generate the associated caption. We use the same region features as VinVL [29], which are released on their project page. Since the object detection model Omni-detection used in previous works [13, 29] is not available, we replace it with a publicly available model of TSD [44] to generate the object detection tag.

Reproducing our method. We perform VIVO [13] pre-training for 100 epochs with a batch size of 1024 and a learning rate of 5×10^{-5}, which are exactly the same as the parameters stated in the VIVO paper. After that, we propose to train our model following the training process described in Algorithm 1 to learn to caption novel objects with linguistic fluency. We train our model for 20 epochs with an effective batch size of 512 (256 caption-labeled images and 256 uncaptioned images) and a learning rate of 1.5×10^{-5}. Then, to learn novel object captions with fidelity and adequacy, we train our model as decsribed in Algorithm 2. Specifically, we train our model for 4 epochs with an effective batch size of 128 (64 caption-labeled images and 64 uncaptioned images) and a learning rate of 2.5×10^{-6}. We use 8 V100 GPUs to perform the above training algorithms. Codes can be found in the supplementary materials.

Reproducing baseline methods. For VinVL [29], we leverage the released model on their project page and directly inference on the nocaps dataset. However, for VinVL+VIVO [29], since the pre-trained model is not publicly available, we reproduce this method using the image region features and object detection tags generated by models mentioned in the beginning of this section to train this model. Specifically, the model is trained for 160K iterations (about 100 epochs) with a batch size of 1024 and a learning rate of 5×10^{-5}, and fine-tuned for 30 epochs with a batch size of 256 and a learning rate of 5×10^{-5} using the cross-entropy loss. Last, we perform the SCST optimization [14] with a learning rate of 2×10^{-6} for 5 epochs to obtain the final model. The numbers reported in Table 7 are derived using this version of model.

Table 8: Image captioning evaluation results on COCO "Karpathy" test split. Note that B@4 stands for BLEU@4, M for METEOR, R for ROUGE-L, C for CIDEr, and S for SPICE.

	B@4	M	R	C	S
VinVL	39.8	29.9	59.6	134.6	23.9
VinVL+VIVO	39.7	29.9	59.6	134.5	23.8
Ours	**40.0**	**30.4**	**60.2**	**137.3**	**24.5**

Table 9: Ablation studies of the joint-training model on nocaps validation set.

Method	in-domain		near-domain		out-of-domain		overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
Baseline (Only w/ L_{s2s})	96.1	13.71	90.35	13.41	79.96	11.77	89.07	13.13
+ L_P	99.44	13.91	91.13	13.53	81.11	11.82	90.29	13.25
+ r_{CIDEr}	109.14	14.52	100.66	14.08	88.61	12.69	99.43	13.87
+ r_{A}	103.81	**15.99**	98.91	**15.32**	93.17	**13.67**	98.45	**15.09**
+ r_{rep} (Ours)	**110.56**	15.23	**105.16**	14.81	**96.22**	13.19	**104.12**	14.55

C Additional experiments

C.1 Detailed ablation analysis

Table 7 lists the performances and compares contributions of the imposed objectives in our P2C. The baseline model in Table 7 is only trained on the COCO Caption dataset using the sequence-to-sequence objective. To confirm our introduction of exploiting BERT to learn the associated wordings of novel images, we apply L_P to the baseline model, and report the results in the second row of Table 7. The CIDEr scores improve significantly after adopting reinforce algorithm [33] and using CIDEr scores of the generated captions as reward, and the results are in the third row. One can see that the SPICE scores largely increase but the CIDEr scores slightly decrease after the deployment of the association model A. We hypothesize that the captioning model properly captures the visual content in images, but it describes the scene with poor linguistic fluency. As the discussion in Sec. 3.2, we attribute the performance drop to the degenerate solution of increasing the association between the captions and the corresponding images. Note that we further consider the repetition penalty to regularize the captioning model. The results are shown in the last row of Table 7. One can see that this regularization slightly improves the SPICE scores but significantly increase the CIDEr scores. By comparing the performances listed in Table 7, we see that the full version of our P2C achieved the best performance in terms of CIDEr and SPICE. Thus, the design of our P2C can be successfully verified.

C.2 Experiments on the COCO Caption dataset

To validate that our method generalize well on the task of describing the seen objects, we conduct experiments on the COCO Caption test set and report the numbers in Table 8. The training data for VinVL [29] is image caption pairs from the COCO [45] dataset. While for VinVL + VIVO and our method, we additionally leverage the uncaptioned image from the Open Images [37] dataset as extra data. One can see that our method outperforms the other competitive approaches on different metrics which verifies the effectiveness of our approach.

C.3 Experiments on the nocaps (XD) benchmark

Recall that in Sec. 4.2, we quantitatively show that our method surpasses current state-of-the-art large-scale methods even if we use a smaller training corpus. In the subsection, we would like to investigate whether the improvement from our module design is still significant when we scale up training data.

To quantitatively show that the performance gain in Table 2 is not simply contributed by the additional data we considered, we ablate our model on the nocaps validation set and show the results in Table 9. We observed a similar performance trend as we reported in Table 7, where L_P slightly improves

Table 10: Human study on the nocaps validation set.

Method	M1 (Turing Test)	M2 (Fluency)	M3 (Fidelity)	M4 (Adequacy)
VinVL +VIVO	0.25	3.99	3.70	3.46
Ours	0.43	4.06	4.33	**4.24**
Human	**0.53**	**4.09**	**4.44**	4.18

the CIDEr scores, and r_{A} significantly boost SPICE but slightly deteriorates the CIDEr scores. One can see that the regularization r_{rep} slightly improves the SPICE scores but significantly increase the CIDEr. By comparing the performances listed in Table 7 and Table 9, we see that our design of using a paraphrase model P to enhance fluency (in terms of CIDEr) and the uses of the association model A to encourage captions with sufficient fidelity and adequacy (in terms of SPICE) still function properly when more diverse image-caption pairs are considered, verifying the design of our P2C.

C.4 Human study

To conduct human study, we randomly picked 60 images from the nocaps validation set, and compared the captions generated by our method to those generated by the SOTA of VinVL+VIVO [29], and the human-annotated captions provided by the nocaps dataset. Following the evaluation protocols used in the COCO Captioning Challenge 2015 [45], we designed 4 different metrics and asked individuals to evaluate captions from these aspects. The following are the four metrics we used in the experiment: M1: Is the caption generated by human (0: machine, 1: human)? (Percentage of captions that pass the Turing Test.) M2: Rate the correctness of the captions on a scale 1-5 (incorrect-correct): Whether the described objects or activities are correct. M3: Rate the amount of detail of the captions on a scale 1-5 (lack of details - very detailed): Whether the caption has detailed all the objects and their attributes. M4: Rate the fluency of the captions on a scale 1-5 (lack of fluency-very fluent): Whether the caption use phrases/words that human generally would use to describe the scene, i.e., the caption is linguistically natural and fluent.

Specifically, M2, M3, M4 correspond to the fidelity, adequacy, and fluency, respectively, which are the particular objectives desired to be achieved. We asked 24 people two answer 6 different questionnaires, and each questionnaire contains 10 captions from each method (i.e., ours, sota, and human caption presented in a random order). We report the results in Table 10. We see that our method surpassed the SOTA by clear margins, while our performances were comparable to those the human ones across different metrics. This further supports the design of our model for NOC with sufficient fluency, fidelity, and adequacy.

C.5 More qualitative results

Qualitative comparison on fluency, fidelity and adequacy. In this part, we provide more qualitative results on the nocaps validation/test set, and the results are shown in Fig. 4 and 5. Note that wordings that are less accurate or incorrectly describe the associated visual content are marked in bold. And, our wording improvements are highlighted in red. Take results in the bottom row of Fig. 4 for example. For the column of fluency, our model particularly described the turtle being *"crawling on some rocks"* instead of *"sitting on the top of a beach"*. For fidelity, our model predicted the background preferably as *"race track"* instead of *"street"* from the prediction of VinVL model. As for the column of adequacy, though both captions described a young men running, our model successfully captures more details in the image (i.e., *"there are number on their shirts"*). For more qualitative results, please refer to Fig. 5.

Qualitative results of some failure cases. In this part, we demonstrate some failure cases of our P2C model. We empirically observe that the failure cases mainly come from the wrong/missing detection tags predicted by the pre-trained object detectors. To be more specific, the captioning model largely relies on the detection tags as clues to correctly describe novel objects. Take the result in the left-side of Fig. 6 for example, the detection model falsely recognizes the raccoon as a squirrel, and this detection result consequently damages the caption prediction. Therefore, how to jointly improve

		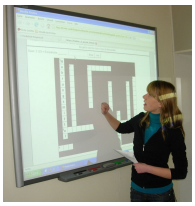	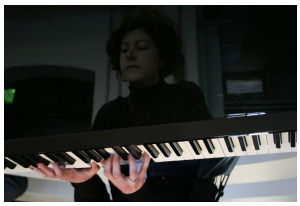
	Fluency	**Fidelity**	**Adequacy**
VinVL	a group of men **standing** in a field with watermelon.	a woman standing in front of a **computer screen.**	a woman sitting on a piano.
Ours	a group of men **cutting up** watermelon in a field.	a woman standing in front of a **projector screen** with a presentation.	a woman sitting in front of a piano **playing a keyboard**.
	Fluency	**Fidelity**	**Adequacy**
VinVL	a tortoiset **sitting** on top of a beach.	a group of men running down a **street**.	a group of young men running in a field.
Ours	a large turtle **crawling** on some rocks in the dirt.	a group of men running down a **race track**.	a group of young men running in the grass **with numbers on their shirts**.

Figure 4: Example results and comparisons for image captions produced by VinVL and ours in terms of fluency, fidelity and adequacy. Note that both utilize VIVO for novel object detection.

the detection model and captioning model is still a open question, and we leave this problem for future research. For more failure cases, please refer to Fig. 6.

	Fluency	Fidelity	Adequacy
VinVL	a couple of people **sitting** on a red bike.	a **black vase** sitting on top of a **table**.	a man riding a pink bike in the street.
Ours	a couple of people **riding** on a red bike.	a **yellow lamp** with a light bulb on a **black background**.	a man riding a bike on a street **with a helmet**.

	Fluency	Fidelity	Adequacy
VinVL	a yellow bee **sitting** on top of blue flowers.	a woman wearing a hat and **a table**.	a woman holding a cell phone in her hand.
Ours	a yellow bee **flying** next to a bunch of purple flowers.	a woman wearing a **brown hat and smiling.**	a woman **wearing a brown jacket and boots** holding a cell phone.

Figure 5: Example results and comparisons for image captions produced by VinVL and ours in terms of fluency, fidelity and adequacy. Note that both utilize VIVO for novel object detection.

Detection tags	Tree, **Monkey**, **squirrel**	Man, **Football** helmet, Sports uniform, **Baseball glove, Baseball bat**	**Dog**, Carnivore, Lion, Brown bear
GT tags	**Raccoon**	**lacrosse stick**	**Jaguar**
Ours	a brown **squirrel** sitting on a tree branch.	a man holding a **tennis racket** on a field	a collage of pictures of **dogs** and a lion.

Figure 6: False captions misled by the wrong object detection tags.