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);
  $\hat{y}_u^p \leftarrow \text{P}(\hat{y}_u^m)$
  $\mathcal{L}_{s2s} \leftarrow \text{CrossEntropy}(\hat{y}_l, y_l)$
  if $A(x_u, \hat{y}_u^b) \leq A(x_u, \hat{y}_u^c)$ **then**
    $\mathcal{L}_P \leftarrow 0$
  else
    $\mathcal{L}_P \leftarrow \text{CrossEntropy}(\hat{y}_u^c, \hat{y}_u^p)$
  end
  $\mathcal{L} \leftarrow \mathcal{L}_{s2s} + \mathcal{L}_P$
  Update parameters: $\theta \leftarrow \mathit{Adam}(\theta, \eta_{it}, \nabla_\theta \mathcal{L})$
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 $\tilde{\text{y}}$ and the associated wordings as $\hat{\text{w}}$, we define fluency as the probability of the NOC model which correctly predicts the collocation given the novel image context $p(\hat{\text{w}}|\tilde{\text{y}})$ (N as the number of collocations). In natural language processing, masked language models are widely applied to predict the masked word $\hat{\text{w}}$ given the context $\tilde{\text{y}}$. Thus, the objective of a masked language model would be maximizing the log-likelihood of the masked word $\log(p(\hat{\text{w}}|\tilde{\text{y}}))$ given the context $\tilde{\text{y}}$, 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: $\theta \leftarrow \textit{Adam}(\theta,\ \eta_{it},\ \nabla_\theta \mathcal{L}_{RL})$
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 publicily 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

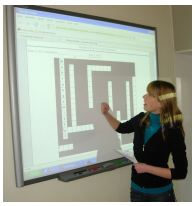
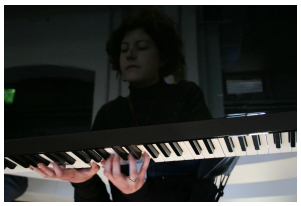
To conduct human study, we randomly picked 60 images from the nocaps validation set, and compared the captions generated by our method to those generated by the SOTA of VinVL+VIVO [29], and the human-annotated captions provided by the nocaps dataset. Following the evaluation protocols used in the COCO 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**.



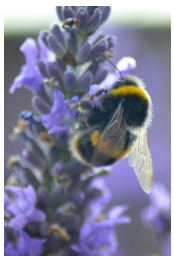
	Fluency	Fidelity	Adequacy
VinVL	a tortoiset **sitting** on top of a beach.	a group of men running down a **street**.	a group of young men running in a field.
Ours	a large turtle **crawling** on some rocks in the dirt.	a group of men running down a **race track**.	a group of young men running in the grass **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.



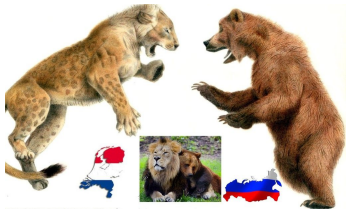
	Fluency	Fidelity	Adequacy
VinVL	a couple of people **sitting** on a red bike.	a **black vase** sitting on top of a **table**.	a man riding a pink bike in the street.
Ours	a couple of people **riding** on a red bike.	a **yellow lamp** with a light bulb on a **black background**.	a man riding a bike on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