We thank all the reviewers for their helpful comments and for recognizing the novelty of our approach (R2–4) and its potential for positive impact (R2, R4). We are glad that the reviewers found our experimental setup exhaustive (R1–4).

Technical contribution (R1, R3). The novelty of our approach lies in the manner in which we formulate our COS latent space in Eq. (3) to encode the object and context information, allowing to better capture the underlying distribution. This enables the exploitation of contextual similarities between images and captions beyond that from paired training data in our pseudo-supervised setup. This is not feasible with prior work, *e.g*. Seq-CVAE [6]. This is the first work that addresses diversity *and* novel object captioning in a single framework (Sec. 3.2, Eq. (7), ll. 198–212).

Oracle evaluation and consensus re-ranking (R2). The oracle evaluation has been used extensively in prior work on multi-prediction (Guzmán-Rivera et al., NIPS 2012). The oracle evaluation does score the top sentence, but is evaluated for a limited number of sentences per image and averaged across the test set. Thus, for a high oracle score the model needs to put high probability mass on the likely captions for an image without relying on "chance". The oracle score provides an upper bound to any re-ranking method. While consensus re-ranking can indeed be improved using stronger networks to close the gap between oracle performance and consensus re-ranking, we follow the same consensus re-ranking procedure as [6,14] for fairness and achieve much better accuracy (Tab. 9, suppl.). Furthermore, the accuracy after re-ranking (Fig. 4) is competitive to methods that generate a single caption, *e.g*. [4], using beam search. We will move Tab. 9 (suppl.) to the main paper. We compute $p(x|z, I)$ and, therefore, ranking captions x for an image I by decoder probabilities $p(x|I)$ as in beam search would require marginalization over z, which is computationally difficult.

Accuracy scores (R2, R3, R4). Consensus re-ranking provides a realistic evaluation against the competing methods (Tab. 9). To further validate our approach, we compare the best-10 sentences after consensus re-ranking in Fig. 4 against Seq-CVAE [6] and beam search on a standard captioning method of [4], which has the same base network. The accuracy results show that COS-CVAE is competitive to beam search and thus diverse *yet* accurate. The average(worst) CIDEr for all the 20 samples with COS-CVAE is 0.875(0.457) compared to 0.792(0.351) for Seq-CVAE. We will add this.

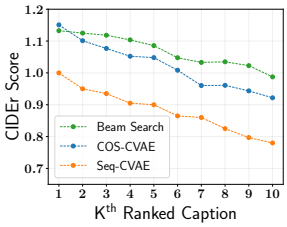

Figure 4: CIDEr score of consensus re-ranked best-10 captions from 20 samples.

F1-score, Fig. 3b (R2, R4). Ranking captions by decoder probabilities is not feasible as discussed above. Consensus re-ranking is not possible as the training data does not contain novel objects. The test data statistics for held-out COCO [17] show that ∼57% of the images with novel objects have at least one ground-truth caption *without* the mention of novel object. For ground-truth annotations, mention of novel object in any of the sentences is counted towards the F1-score [3]. Therefore, it would be limiting to expect every caption from a diverse captioning approach, which learns the distribution of the likely captions, to mention the novel object (Tab. 11). We will clarify and highlight these challenges in the final version. We thus show the F1-score and the highest CIDEr for sample sizes of z in $\{5, \ldots, 100\}$ in Fig. 3b and demonstrate competitive results even for a small sample size of 5. Our COS-CVAE with 5 samples has the highest CIDEr of 1.111 and an average CIDEr of 0.775, which is competitive to NBT [30] with a similar base network. We will add this.

Effect of context-object split (R2). Verbs, adjectives, and spatial relationships are crucial components of the context (ll. 105). Random samples in Tab. 2, 11, and 12 show that the captions from COS-CVAE are coherent (ll. 287–288).

Self-CIDEr (R3). COS-CVAE has a score of 0.742 while Seq-CVAE(attn) has 0.714. We will add a human evaluation.

Diversity metrics, Tab. 4 (R4). Consistent with prior work, *e.g*. [6], Unique and Novel are computed with 100 samples and Div-1, Div-2, and m-Bleu are computed with top-5 generated captions after consensus re-ranking (ll. 251–253).

Benefit of pseudo supervision (R3). Pseudo supervision allows to leverage additional contextual descriptions from human annotations that can describe the image, thus capturing the true diversity of the underlying distribution (ll. 94–96). We use pseudo supervision along with ground-truth captions for best performance because ground-truth captions help accurately guide the attention of the decoder (ll. 128–130). The application of pseudo supervision to a retrieval task is an interesting idea and can be considered for future work since the extension is non-trivial.

CBS constraints (R4). Seq-CVAE (attn)+CBS yields an F1-score of 55.4 with 5 samples compared to an F1-score of 70.1 with COS-CVAE+CBS. However, CBS decoding leads to less diverse captions (one of the main goals here).

Clarifications (R2). The results of Div-BS are on the same test split as [6, 14] (*cf*. Tab. 1 in [6]). The base vision network for different methods can be different as various models impose different constraints in the latent space [14, 41]. We have included the Seq-CVAE(attn) baseline with the same base vision network as our COS-CVAE.

Nocaps dataset (R1, R2, R4). Our variational COS-CVAE framework models the distribution of likely captions for a given image. The evaluation server for the nocaps dataset (Agrawal et al.), however, accepts only one caption per image and does not support methods modeling one-to-many relationships for images and captions. We thus provide the results on the real world held-out COCO dataset for novel object captioning [17] instead.

Additional references (R2, R3). Thank you. We will add this.