We thank all the reviewers for the helpful comments. We will revise the paper to address your concerns.

R1-Q1: The implementation seems straight-forward and the ablation analysis on the loss function.

In the beginning, based on the Up-Down model, we have attempted to implement the Constant Prophet Attention with more complex models, i.e., single-head dot-product attention (120.2 CIDEr), multi-head dot-product attention (118.8 CIDEr), bilinear attention (122.8 CIDEr) and attention-based LSTM (121.7 CIDEr). However, they all lower the performance of Up-Down base model (123.5 CIDEr on MSCOCO) and introduce extra model parameters and inference computations. While our current implementation can boost the performance without introducing any additional parameters or slowing down inference computations. For ablation analysis, our preliminary experiments showed that using L1 norm improves the performance. Thus we kept using L1 norm in the rest of experiments. We follow your constructive advice to apply the L2 norm and KL divergence to the Up-Down model. The results with L1 norm, L2 norm and KL divergence are 129.0, 128.2 and 126.9 CIDEr which all outperform the original Up-Down model. It shows that L1 norm achieves the best performance and all loss functions are viable in practice with substantially improved performance. We will conduct a systematic comparison between various loss functions in the next revision.

R2-Q1: The attention for some words should not only from the future, but also from the history.

For noun phrases, our approach uses both history ($i < t$) and future ($j > t$) information (see Eq. (5) and Eq. (9)). For other cases like verbs, we did not consider the explicit history information. However in the process of sequence generation, the hidden state could contain implicit information in previous time stamps. Nevertheless, we will attempt to adopt the scene graph, which is able to extract triples from a sentence or an image. The triples like <*subject-verb-object*>, e.g., <*woman-holding-umbrella*>, can provide more explicit history information.

R2-Q2: The first visualization results of DPA in Figure 3 of our paper.

In this example, the top-1 attended region for the noun phrase "a smiling boy" is '*mouth*'. Therefore, in the context of DPA, the attended region of the word "smiling" is the same as the noun phrase "a smiling boy".

R3-Q1: The value of the hyper-parameter λ.

To choose a proper value of λ, we follow the setting of the regularization losses like weight decay and adversarial loss which choose small λ values. For example, the weight of L2 regularization is "tiny", often set as low as 1e-4, but has a positive effect on the model training. We also evaluate the performance of different settings of λ in Table 4 to ensure that the degree of our regularization is tuned to match the model training.

R3-Q2: The evaluation on non-caption generation tasks and the 1st place on the leaderboard.

Thanks for your great suggestion for the evaluation on non-caption generation tasks. It is a good future direction and we will work on it. For the leaderboard, CIDEr-c40, specially designed for captioning, is the default ranking metric, which is more convincing than CIDEr-c5, as shown in Vedantam et al. [2015] that CIDEr achieves higher correlation with human judgment when more reference sentences are given. It is worth noticing that most current published top rankers use ensemble. However, it's true that we are not aware of whether other submissions use ensemble or not, if they did not publish their approaches. We will tone down our voice in the next revision.

R4-Q1: How much the model slows down?

In the inference stage, our approach does not incur extra computational cost. In the training stage, extra computations are mainly introduced by the calculation of L1 loss (see Eq. (6)) and the Prophet Attention (see Eq. (9)). We find that our approach slows down the training procedure around 4%. We will provide detailed information in the next revision.

R4-Q2: The grammatical and clerical errors. What is the difference between our work and previous works?

Thank you very much! We will polish this paper and fix the grammatical and clerical errors. For these three related works, we will cite and discuss them in detail. Although [1,2,3] attempted to exploit the future information, their modelings are different from ours. Given a time stamp in sequence generation, in addition to current target, [1,2,3] introduce new model parameters to further predict the future words. Specifically, [1,2] further predict target word one more step ahead, and [3] predicts the rest of the caption. The prediction accuracy of future targets and current target are combined together in parameter optimization. While in our proposed approach, we use the hidden states of future time stamps to explicitly guide the attention calculation of current one without introducing additional model parameters or inference computations. In other words, we aim to better ground current target word to proper image regions which alleviates the attention deviation problem, resulting in improved performance of both grounding and captioning.

References

R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.