We authors are grateful to the reviewers for their valuable comments. We will improve the final version by taking all the review comments and release the source code package to ensure the reproducibility. Below, we number and address comments of each reviewer in order.

**R1.1** - Why is the probability of \( w \) matched by \( l \) calculated over the minibatch in Eq (3)?

\[ p(w|l) \]

\[ \hat{p}(w|l) \]

**R1.2** - Why not use a symmetric similarity measure? An exploration of the effect of using a cosine similarity.

\[ s_{w|l} \]

\[ s_{l|w} \]

**R1.3** - The imagination phase is just sampling a noise vector.

**R1.4** - What is the effect of having the adversarial term in the encoder loss? An ablation study of this adversarial term.

**R2.1** - How can the model be able to control each word as presented in the experiments?

**R2.2** - This paper only compares with AttnGAN and ablation studies of key components are missing.

**R2.3** - The paper names the text-image encoded feature as the prior knowledge, which is overclaimed.

**R3.1** - Improving the limitations the authors already discussed in the paper.