We authors are grateful to the reviewers for their valuable comments. We will improve the final version by taken 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.

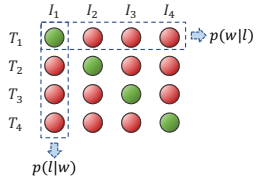

Figure 1: Illustration of matched & mismatched (T_n, I_n) pairs in a minibatch.

Table 1: Results of LeicaGAN trained with E_T^I and E_T^M of different loss weights.

Evaluation Metric	Inception Score		R-precision [%]	
	CUB*	Oxford-102*	CUB*	Oxford-102*
(1) LeicaGAN, $E_T^I(\alpha_1$=0), $E_T^M(\alpha_2$=0)	5.51±0.03	3.64±0.01	82.89	84.02
(2) LeicaGAN, $E_T^I(\alpha_1$=1), $E_T^M(\alpha_2$=0)	5.58±0.05	3.71±0.02	85.20	85.55
(3) LeicaGAN, $E_T^I(\alpha_1$=5), $E_T^M(\alpha_2$=0)	5.60±0.05	3.68±0.01	82.95	84.03
(4) LeicaGAN, $E_T^I(\alpha_1$=1), $E_T^M(\alpha_2$=2)	5.63±0.04	3.79±0.01	84.57	84.98
(5) LeicaGAN, $E_T^I(\alpha_1$=1), $E_T^M(\alpha_2$=4)	**5.69±0.05**	**3.80±0.01**	**85.28**	**85.81**
(6) LeicaGAN, $E_T^I(\alpha_1$=1), $E_T^M(\alpha_2$=6)	5.61±0.06	3.64±0.02	81.65	84.01

R1.1 - Why is the probability of w matched by l calculated over the minibatch in Eq (3)?
⇒ Assuming a minibatch of 4 pairs of (T_n, I_n), where $n \in \{1, 2, 3, 4\}$ as shown in Fig.1, we treat the green and red pairs as matched and mismatched pairs respectively. Therefore, for each text, we obtain 4 similarity scores indicating the closeness between the given text with the matched & mismatched images (vice versa). Afterwards, according to Eq.(4), $p(w|l)$ and $p(l|w)$ are estimated separately. Specifically, a softmax is applied among the matched and mismatched text-image (image-text) pairs in a batch along the row (column) as shown in Fig.1. They serve as different constraints for training the model at the same time.

R1.2 - Why not use a symmetric similarity measure? An exploration of the effect of using a cosine similarity.
⇒ The local-level similarity score $s_{w|l}$ or $s_{l|w}$ of the (T, I) pair is first calculated as: $log(\sum_{i=1}^{L} exp(\gamma_1 cos(\widehat{l_i}, w_i)))^{1/\gamma_1}$ [37], which is a symmetric similarity result and we elaborate this equation here for a more clear explanation of the calculation. Careful revision of this part and further exploration about different similarity measures will be included in the final version.

R1.3 - The imagination phase is just sampling a noise vector.
⇒ The embedded textual features obtained in TVE convey different kinds of visually-relevant information. Therefore, in the imagination phase, an aggregation of these features and noise is applied to obtain an initial visual impression serving as the input for the subsequent creation phase. The noise is added to guarantee the diversity of the image generation. In the final version, we will explore more efficient and diverse modules to strengthen the impact of the imagination phase as we discussed in the section of Limitation.

R1.4 - What is the effect of having the adversarial term in the encoder loss? An ablation study of this adversarial term.
⇒ The loss weights α_1 and α_2 in training text-visual co-embedding models E_T^I and E_T^M, respectively, were carefully studied in Table 1. The detailed results will be reported in the final version. The comparison between (1) and (2) shows that the use of the adversarial loss obtains a performance gain. An intuitive illustration of the effect of the adversarial loss reducing the multi-modal domain gap has been shown in Fig.4 of [35].

R2.1 - How can the model be able to control each word as presented in the experiments?
⇒ The input sentence was first embedded into both (1) word-level feature matrices w to represent the semantic meaning of the words and (2) a sentence-level feature vector s to convey the semantic of the whole sentence. Afterwards, through the co-embedding of images and text, the correlations between them can be built. Therefore, during the decoding stage, with the guidance of the attention module, the generators focused on different words, resulting in different visual outputs accordingly.

R2.2 - This paper only compares with AttnGAN and ablation studies of key components are missing.
⇒ Before the NeurIPS submission deadline, StackGAN and AttnGAN were recognized as state-of-the-art models in T2I. Recently, Obj-GANs, SD-GAN, MirrorGAN, and DM-GAN are published in CVPR2019. However, it is not straightforward to compare LeicaGAN with all of them because they were tested on different datasets using different evaluation metrics and only two of them (MirrorGAN and Obj-GAN) released their source code packages in the last two months, after the NeurIPS submission deadline. To make a fair comparison, we will re-train and test these models under the same condition in the final version. Based on the results in Table 2 in the paper, we verified the effectiveness of the following key components: (1) the global-local attentive generator, (2) E_T^M, and (3) the collaborative attention module and its weights. Here, we report the ablation study for the weight choosing of the TVE models in Table 1. More details will be provided in the appendix and code.

R2.3 - The paper names the text-image encoded feature as the prior knowledge, which is overclaimed.
⇒ We name the text-image encoded features as the prior knowledge because the co-embedding mappings, which correlate textual and visual features with each other, are learned in the PKL phase. These cross-modal embeddings serve as the prior knowledge to guide the image generation of the subsequent phases. We can expect that this phase can be further improved by incorporating different types of prior. We agree the name of "prior knowledge" can be inappropriate and will change it to "prior" in the final version.

R3.1 - Improving the limitations the authors already discussed in the paper.
⇒ In the future, we plan to introduce more prior to the first PKL phase, such as fine-grained attributes, and more efficient and diverse modules will be explored to strengthen the impact of imagination. Additionally, more advanced network structures and training strategies will be employed in order to further enhance the generation ability of the proposed T2I model.