A text to image generation (T2I) model aims to gener#ate photo-realistic images which are semantically consis#tent with the text descriptions. Built upon the recent ad#vances in generative adversarial networks (GANs), existing T2I models have made great progress. However, a close in#spection of their generated images reveals two major limi#tations: (1) The condition batch normalization methods are applied on the whole image feature maps equally, ignor#ing the local semantics; (2) The text encoder is fixed during training, which should be trained with the image generator