SpaText: Spatio-Textual Representation for Controllable Image Generation

Abstract:

Recent text-to-image diffusion models are able to generate convincing results of unprecedented quality. However, it is nearly impossible to control the shapes of different regions/objects or their layout in a fine-grained fashion. Previous attempts to provide such controls were hindered by their reliance on a fixed set of labels. To this end, we present SpaText - a new method for text-to-image generation using open-vocabulary scene control. In addition to a global text prompt that describes the entire scene, the user provides a segmentation map where each region of interest is annotated by a free-form natural language description. Due to lack of large-scale datasets that have a detailed textual description for each region in the image, we choose to leverage the current large-scale text-to-image datasets and base our approach on a novel CLIP-based spatio-textual representation, and show its effectiveness on two state-of-the-art diffusion models: pixel-based and latent-based. In addition, we show how to extend the classifier-free guidance method in diffusion models to the multi-conditional case and present an alternative accelerated inference algorithm. Finally, we offer several automatic evaluation metrics and use them, in addition to FID scores and a user study, to evaluate our method and show that it achieves state-of-the-art results on image generation with free-form textual scene control.