Prompt-Free Diffusion: Taking “Text” out of Text-to-Image Diffusion Models – Supplementary –

Xingqian Xu[1,5], Jiayi Guo[1,2], Zhangyang Wang[3,5], Gao Huang[2], Irfan Essa[4], Humphrey Shi[1,4,5]
[1]SHI Labs @ UIUC, Georgia Tech & Oregon, [2]Tsinghua University, [3]UT Austin, [4]Georgia Tech, [5]Picsart AI Research (PAIR)
https://github.com/SHI-Labs/Prompt-Free-Diffusion

1. Quantitative Measurement

We conducted the following experiments on the benchmark dataset COCO Caption, measuring the performance with FID and CLIP scores.

Method	**SeeCoder**	**CNet**	**FID** (↓)	**CLIP-I** (↑)
SD15	-	-	14.30	66.44
RV20	-	-	14.98	67.59
RV20	-	Shuffle	14.82	73.57
PFD-PA	✓(PA)	-	12.09	79.88
PFD-Canny	✓	**Canny**	**9.05**	**83.48**
PFD-Depth	✓	Depth	9.31	81.83

Here we would like to explain more details: Data-wise, we use a subset of COCO Val containing 10000 samples and keep them across all tests. We conduct two baseline experiments with SD15 and RealisticVision20 (RV20), generate images using ground truth captions, and compute CLIP-Similarity and FID against ground truth images. We then conduct an experiment with ControlNet (CNet) shuffle (our competitor solution), in which we ignore captions and use ground truth images as conditioning inputs. Similarly, the FID and CLIP scores are computed between ground truth and generated images. Notice that CNet-Shuffle increases the CLIP score from 67.59 to 73.57, revealing the effectiveness of using reference image conditionings compared with text conditionings. We then test three settings of our Prompt-Free Diffusion (PFD): the position-aware (PA) version that requires no structure inputs, and two ordinary versions with canny and depth. For fair comparisons, all PFDs use RV20 as its diffusion UNet. As the table shows, PFDs quantitatively outperform baselines and prior works with noticeable margins, demonstrating the effectiveness of our PFD solution. Among these, PFD-Canny gives the best FID (9.05) and CLIP (83.48).

In addition to the scores shown above, we would like to highlight that PFD can be further applied in a) zero-shot content transfer and b) structural-based and content-based customization and editing. However, these merits cannot be easily quantified under the current testing scheme, which remains an active research area in the community.