On the Scalability of Diffusion-based Text-to-Image Generation

Supplementary Material

A. Evaluation Details

Prompts We generate images with two prompt sets for evaluation: 1) 4081 prompts from TIFA [19] benchmark. The benchmark contains questions about 4,550 distinct elements in 12 categories, including *object*, *animal/human*, *attribute*, *activity*, *spatial*, *location*, *color*, *counting*, *food*, *material*, *shape*, and *other*. 2) randomly sampled 10K prompts from MSCOCO [26] 2014 validation set.

Metrics In addition to previously introduced TIFA [19] and ImageReward [40] scores, we also calculate the following metrics:

- **FID**: FID measures the fidelity or similarity of the generated images to the groundtruth images. The score is calculated based on the MSCOCO-10K prompts and their corresponding images. We resize the groundtruth images to the same resolution (256×256 or 512×512) as the generated images.
- **CLIP**: The CLIP score [14, 32] measures how the generated image aligns with the prompt. Specifically, the cosine similarity between the CLIP embeddings of the prompt and the generated image. Here we calculate it with the MSCOCO-10K prompts and report the average value.
- **Human Preference Score (HPS)** [39]: HPSv2 is a preference prediction model trained with human preference. We calculate the scores based on the TIFA prompts and report the average value.

Inference Settings Given a prompt set and a pre-trained model, we generate images at 256×256 resolution with DDIM [37] 50 steps, using default CFG 7.5 and fixed seed for all prompts. For each model checkpoint, we use its non-EMA weights for evaluation.

B. More Results on UNet Scaling

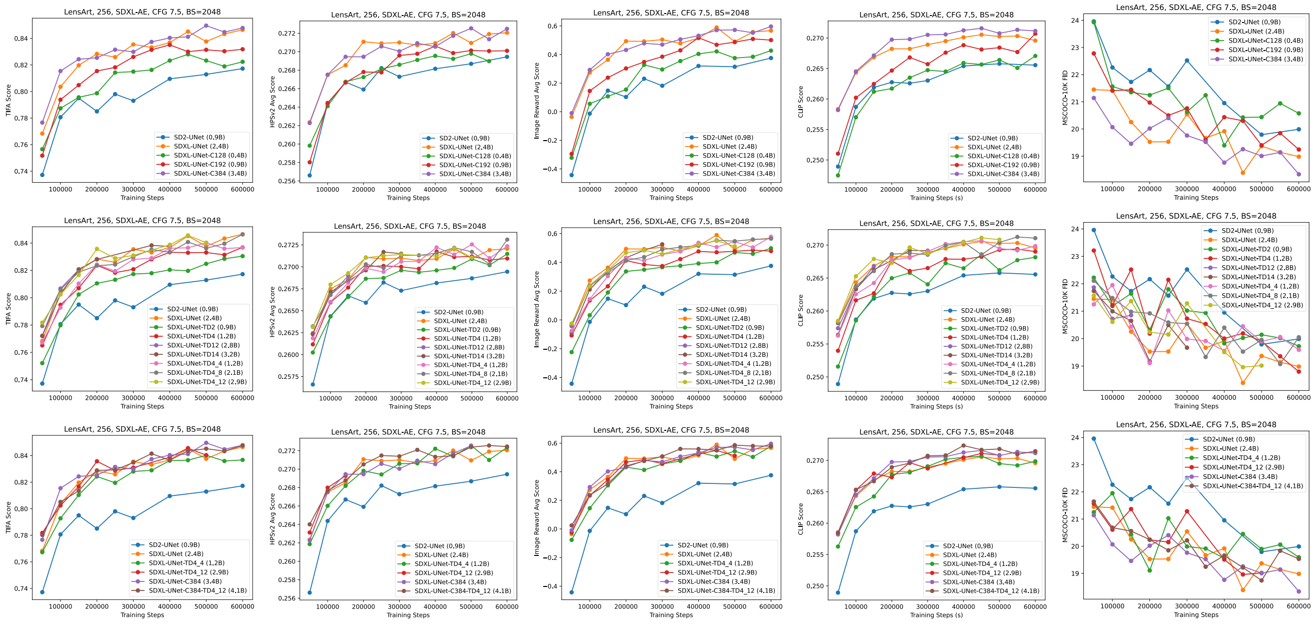

Figure 12. The evolution of all metrics during training for UNet variants. The baseline models are the UNets of SD2 and SDXL. All models are trained with SDXL VAE at 256×256 resolution. The 1st row shows SDXL UNets with different initial channels. The 2nd row shows SDXL UNets with different TDs. The 3rd row compares SDXL UNets with both increased channels and TDs.

Evolution of all metrics for UNet variants We have shown the TIFA evolution curves of SDXL [31] UNet variants in Sec. 3. Here we show the evolution of other metrics during training for all UNet variants in Fig. 12, including the change of *channels*, *transformer depth* and both of them. The pattern of other metrics is very similar as TIFA and the relative performance among models is stable across metrics, e.g., the 1st row of Fig. 12 shows that UNets with more channels tend to have better TIFA, HPSv2, ImageReward, CLIP, and FID scores. Though FID score has more variations during training.

Comparing the training efficiency of SDXL UNet and its variant Previously we introduce a smaller SDXL UNet variant, i.e., TD4_4, which is 45% smaller, 28% faster, and has competitive performance as SDXL-UNet when trained with the same steps (Fig. 12). Here we compare their metrics in terms of training steps as well as the total compute (GFLOPs). We extend the training steps of TD4_4 from 600K to 850K to see whether the performance can be further improved. As shown in Fig. 13, TD4_4 achieves similar or better metrics in comparison with SDXL UNet with much less computation cost. It suggests that TD4_4 is a more compute efficient model when the training budget is limited.

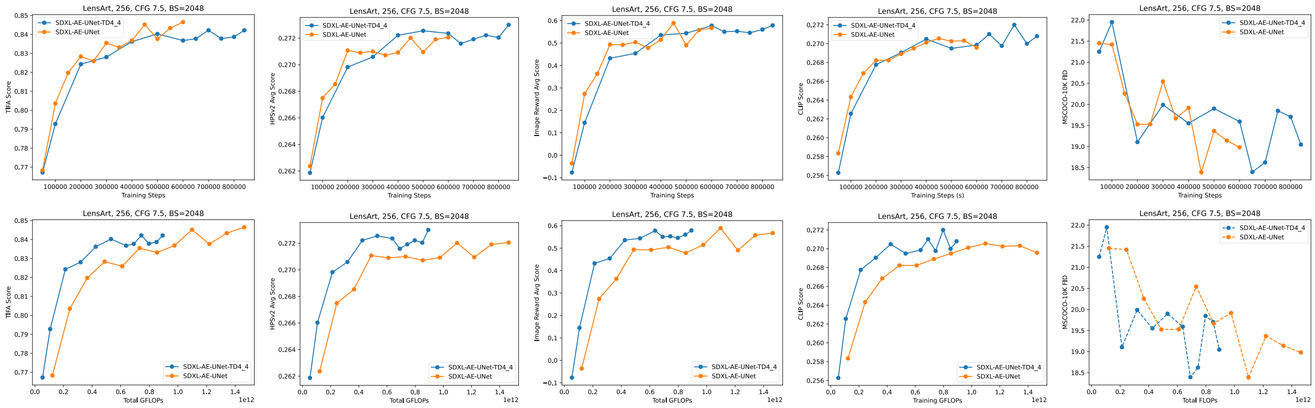

Figure 13. Comparing metrics evolution speed of SDXL UNet and its TD4_4 variant in terms of training steps and total compute (GFLOPs). TD4_4 achieves similar or better metric scores at much less training cost.

C. More Results on Dataset Scaling

Evolution of all metrics for SD2-UNet trained on different datasets We have shown the TIFA and ImageReward evolution curves of SD2-UNet trained on different datasets in Sec. 4. Here we show the evolution of all metrics in Fig. 14. The trend of other metrics is similar as TIFA, except the HPSv2 and CLIP scores for *LensArt-Raw*, which have higher values than *LensArt*. We find the reason is that the *LensArt-Raw* model tend to generate images with more meme text due to a large amount of data has such patterns, and such images usually results in higher values on those two metrics. Those metrics become more precise and meaningful after the training data is filtered by removing those meme images.

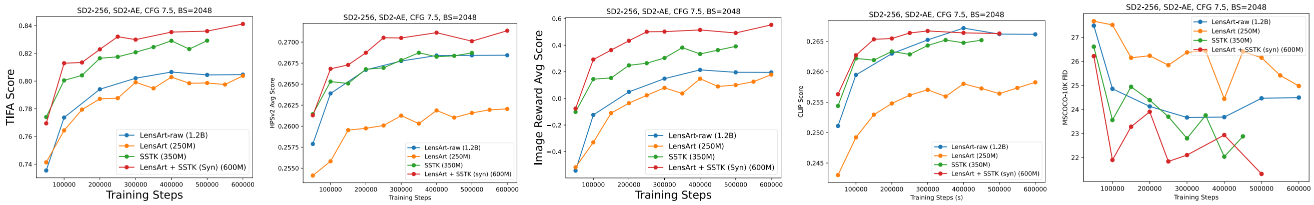

Figure 14. Training SD2 model with different datasets. All metrics show that LensArt + SSTK has better scores than LensArt or SSTK only. Note that the HPSv2 and CLIP scores for LensArt-Raw are much higher than LensArt. The reason is that unfiltered dataset tends to generate images with more meme text.

D. The Effect of VAE Improvement

SDXL [31] introduced a better trained VAE and shows improved reconstruction metrics in comparison with its SD2 version. However, the impacts on the evaluation metrics are not fully explored. Here we ablate the effect of VAE on the evaluation metrics. We compare the training of same SD2-UNet with different VAEs, i.e., SD2's VAE and SDXL's VAE, while keeping all other settings the same. Fig. 15 shows that the improvement of SDXL's VAE over SD2's VAE is significant for all metrics.

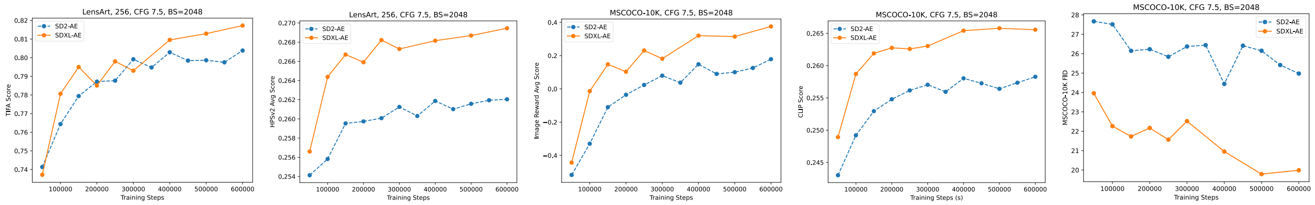

Figure 15. Training SD2 UNet model with different VAEs. The SDXL's VAE has significant improvement on all metrics over SD2's VAE.

E. Scaling the Batch Size

To scale out the training of large diffusion models with more machines, increasing batch size is usually an effective approach. We have been using consistent batch size 2048 in all experiments for controlled studies. Here we also show the effect of batch size on the evolution of metrics. We compare the training of SDXL UNet with 128 channels in different batch sizes, i.e., 2048 and 4096, while keeping other training configs the same. Fig. 16 shows that larger batch size yields better metrics in terms of same iteration numbers. The convergence curve of FID score is more smooth than smaller batch size.

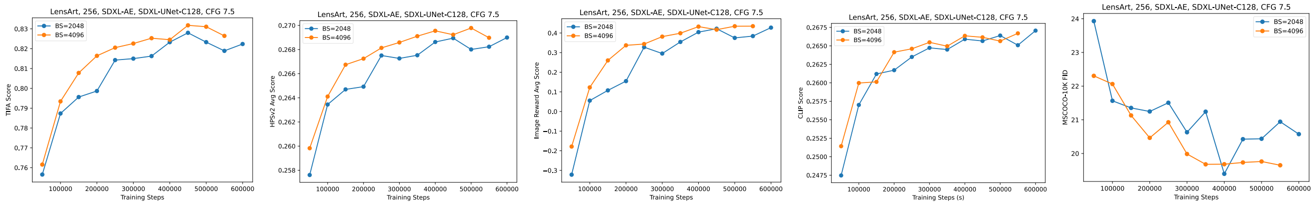

Figure 16. Training SDXL-UNet-C128 with different batch sizes.

F. Model Evaluation at Low Resolution Training

The evaluation metrics at 256 resolution can provide early signals on their performance at high resolutions, which is informative for quick model ablation and selection. The reason is that the high resolution training usually utilizes a subset of images of the dataset, and the text-image alignment and image quality scores usually do not change significantly once they are fully trained at lower resolution, especially the text-image alignment performance. Given two well trained SDXL models (C128 and C192) at 256 resolution, which has clear performance gap, we continue training them at 512 resolution and measure their performance gap. As shown in Fig. 17, both two SDXL UNet models can get performance improvement at 512 resolution, but C128 model still yields worse performance than C192.

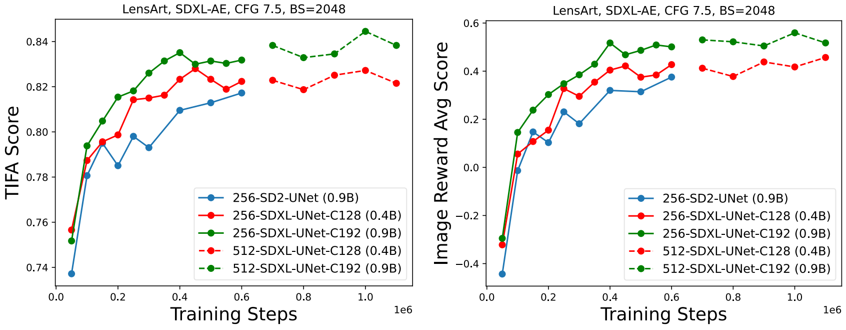

Figure 17. TIFA and ImageReward do not change much during high resolution fine-tuning stage (dashed lines)

G. Caption Analysis

For both LensArt and SSTK dataset, we present the histograms of number of words and nouns of original and synthetic captions respectively in Fig. 18. Note that we overload the noun with noun and proper noun combined for simplicity. First, as shown in the first two figures, we see that synthetic captions are longer than original captions in terms of words, indicating augmenting original captions with synthetic captions can increase the supervision per image. Second, from the last two figures, we note that the number of nouns of synthetic captions are less than those in real captions on average. This is mainly caused by synthetic captions have less coverage in proper nouns, indicting the synthetic captions alone are not sufficient to train a generalist text-to-image model.

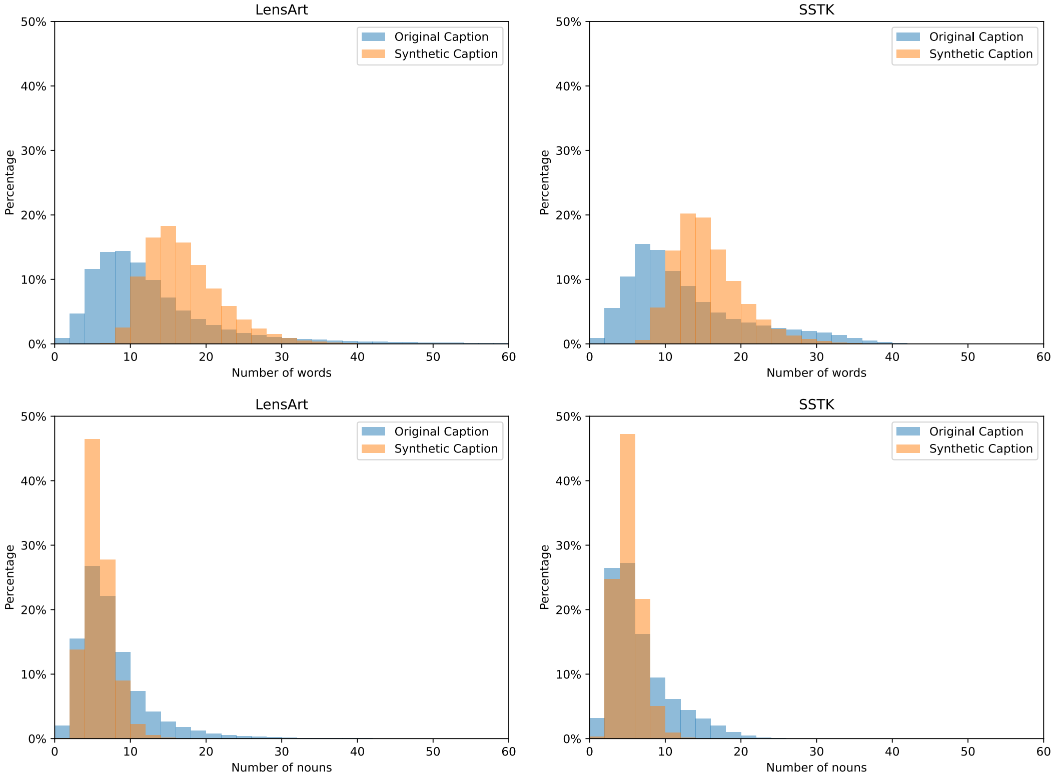

Figure 18. Histograms of word and noun numbers in the original and synthetic captions of different datasets