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

Hao Li[1,2], Yang Zou[1,2], Ying Wang[1,2], Orchid Majumder[1,2], Yusheng Xie[1,2], R. Manmatha[1], Ashwin Swaminathan[1,2], Zhuowen Tu[1], Stefano Ermon[1], Stefano Soatto[1]
[1]AWS AI Labs, [2]Amazon AGI
{haolimax, yanzo, lyiwang, orchid, yushx, manmatha, swashwin, ztu, ermons, soattos}@amazon.com

Abstract

Scaling up model and data size has been quite successful for the evolution of LLMs. However, the scaling law for the diffusion based text-to-image (T2I) models is not fully explored. It is also unclear how to efficiently scale the model for better performance at reduced cost. The different training settings and expensive training cost make a fair model comparison extremely difficult. In this work, we empirically study the scaling properties of diffusion based T2I models by performing extensive and rigours ablations on scaling both denoising backbones and training set, including training scaled UNet and Transformer variants ranging from 0.4B to 4B parameters on datasets upto 600M images. For model scaling, we find the location and amount of cross attention distinguishes the performance of existing UNet designs. And increasing the transformer blocks is more parameter-efficient for improving text-image alignment than increasing channel numbers. We then identify an efficient UNet variant, which is 45% smaller and 28% faster than SDXL's UNet. On the data scaling side, we show the quality and diversity of the training set matters more than simply dataset size. Increasing caption density and diversity improves text-image alignment performance and the learning efficiency. Finally, we provide scaling functions to predict the text-image alignment performance as functions of the scale of model size, compute and dataset size.

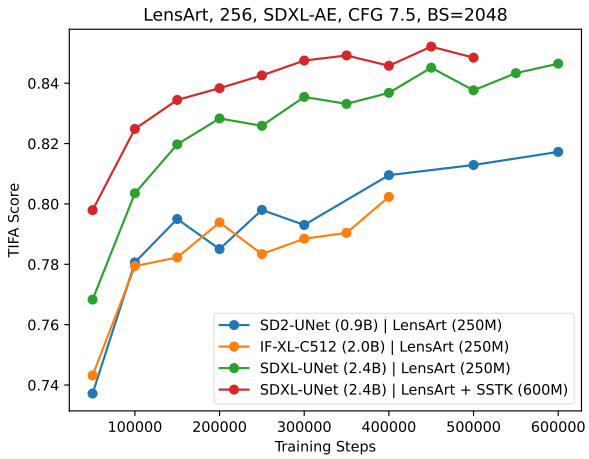

Figure 1. Pushing the Pareto frontier of the text-image alignment learning curve by efficiently scaling up both denoising backbones and training data. Comparing with the baseline SD2 UNet [34], the combined scaling with both SDXL UNet and enlarged dataset significantly increases the performance and speeds up the convergence of TIFA score by 6×.

1. Introduction

Scaling up model and dataset size has been the key enabling factor for the success of LLMs [17, 21] and VLMs [6, 32]. The scaling law [4, 21] governs the expectation of performance as a function of dataset, model size and compute budget. However, the scaling properties for recent diffusion based Text-to-Image (T2I) models [31, 33–35] are not well studied. Though there is emerging trend that T2I models can be improved with larger denoising backbones [9, 31] and stronger text-encoders [1, 31, 35], it is still not clear how to effectively and efficiently scale up diffusion models, e.g., how does the design of denoising backbone influence the image generation and which components are more effective to scale? How should diffusion model scale when the training data increases? To answer the questions, it is essential to understand how exactly each new model improves over previous ones. However, existing diffusion based T2I models are mostly trained with different datasets, input space (latent space or pixel space) and training settings. Moreover, the expensive training cost of high resolution models makes the fair comparison extremely hard, not to mention exploring new ones. Therefore, a fair and controlled comparison of different denoising backbones is greatly desired, which can enable seeking of more efficient models with reduced training and inference cost.

In this paper, we investigate the scaling properties for training diffusion models, especially on the denoising backbone and dataset. The goal is to understand which dimension

of the model is more effective and efficient to scale, how to properly scale the dataset, and the scaling law among models, dataset and compute. Fig.1 gives an illustration of how the Pareto frontier of the text-image alignment performance curve can be pushed via proper scaling.

1.1. What we have done

- **Comparing existing UNets in a controlled environment**: we first compare existing UNet designs from SD2 [34], DeepFloyd [9] and SDXL [31], to understand why certain UNet design is significantly better than others. To allow a fair comparison, we train all models with the same *dataset*, *latent space*, *text encoder* and *training settings*. We monitor multiple evaluation metrics during training, including composition scores and image quality scores. We verified SDXL's UNet achieves superior performance over others with similar amount of parameters, which justifies the importance of architecture design.
- **Scaling UNet and comparing with Transformers**: To understand why SDXL works so well, we conduct extensive ablation studies on the design sapce of UNet by investigating 15 variations ranging from 0.4B to 4B parameters, especially on the choice of channel numbers and transformer depth. We show how each architecture hyperparameter affects the performance and convergence speed. Similarly, we ablate and scale the Transformer backbones [5, 30] and compare with UNet.
- **Ablating the effect of dataset scaling and caption enhancement**: We study how different dataset properties affect the training performance, including dataset size, image quality and caption quality. We curate two large-scale datasets with 250M and 350M images, both are augmented by synthetic captions. We train both small and large models to see how they can benefit from dataset scaling.

1.2. Contributions

- We conduct large-scale controlled experiments to allow fair comparison across various denoising backbones for T2I synthesis, including both UNets and Transformers. Our work verifies the importance of the denoising backbone design. We find composition ability is mainly developed at low resolution, which enables fast model ablations without training in high resolution. To our best knowledge, our work is the first large-scale controlled study allowing fair comparison across different denoising backbones for T2I syntheis.
- We ablate the key design factors for UNet and Transformers and compared their scaled versions. We show scalling the transformer depth in UNet is more parameter efficient in improving the alignment performance in comparison with channel number. We identify an efficient UNet variant that is 45% smaller and 28% faster than SDXL while achieving similar performance. We confirm scaling transformer backbone improves performance, but also identify the difficulty of training from scratch due to lack of inductive bias in comparison with UNets.
- We show that properly scaling training data with synthetic captions improves image quality and speeds up the convergence. We see data scaling can improve small model's performance significantly, a better designed model can have a higher performance upper bound.

2. Related Work

Diffusion Models Diffusion models [15, 16, 28, 29, 36] synthesize samples via an iterative denoising process and have shown superior performance over GAN [13] based methods for image generation [10]. Recent diffusion based T2I models such as Imagen [35], LDM/SD2 [34], DeepFloyd [9], SDXL [31], and DALL·E [3, 33] have shown consistently improved performance in terms of sample diversity, text-image alignment and image fidelity. Pixel-based models [9, 33, 35] usually require cascaded super-resolution (SR) models to upscale images generated in low resolution, while LDMs [3, 31, 34] reduce training cost by utilizing a compressed latent space and upsampling with via an autoencoder [22]. The low resolution latent space may not represent small objects (e.g., faces) well. SDXL mitigates this issue via a better VAE and training models in higher resolution latent space (128×128). Emu [8] shows that increasing the latent channels improves image quality.

Scaling UNets UNet architecture was first introduced for diffusion models in [16]. [10, 28] ablated UNet with several design choices and investigated how FID scales as a function of training compute. The UNet in LDM or SD2 [34] has 320 initial channels and 850M parameters. DeepFloyd [9] trains a pixel based model with a UNet of 4B parameter size and 704 channels, which shows better performance than its smaller versions. SDXL [31] employs a 3× larger UNet than SD2 with mutiple improvements. On the other hand, there are also works on improving UNet's efficiency by scaling it down, e.g., SnapFusion [25] studies the redundancy of UNet and identifies an efficient version by employing the change of CLIP/Latency to measure the impact of architecture change.

Transformer Backbones Recently there is surge interest in using Transformer [38] to replace UNet for its general architecture design and increased scalability [2, 30, 42]. DiT [30] replaces UNet with Transformers for class-conditioned image generation and find there is a strong correlation between the network complexity and sample quality. U-ViT [2] shows comparable performance can be achieved by ViTs with long skip connection. MDT [12] introduces a mask latent modeling scheme to improve the training efficiency of transformer-based diffusion models. Those works

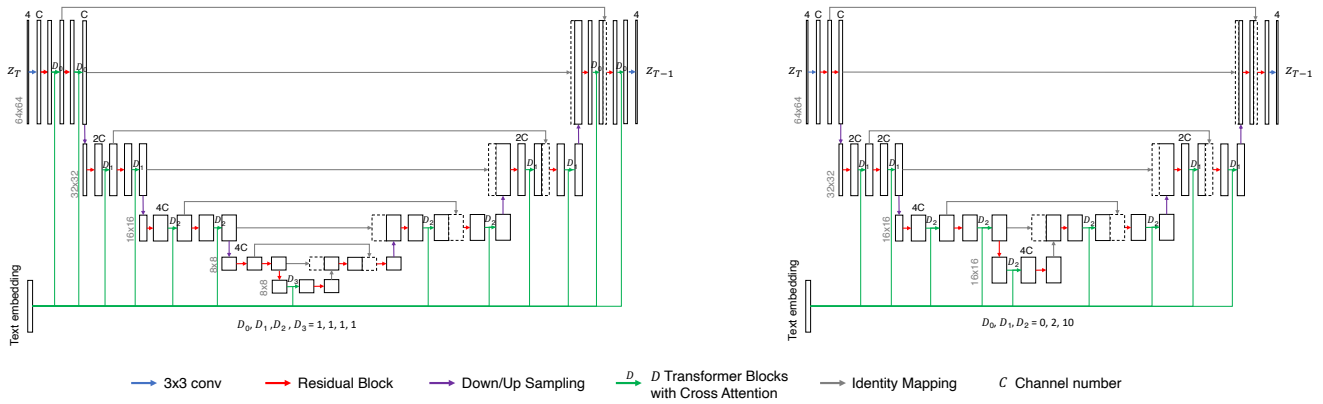

Figure 2. Comparison of the UNet design between SD2 (left) and SDXL (right). SD2 applies cross-attention at all down-sampling levels, including $1\times$, $2\times$, $4\times$ and $8\times$, while SDXL adopts cross-attention only at $2\times$ and $4\times$ down-sampling levels.

are mostly class conditioned models and only the effect of model architecture on image fidelity is studied. PixArt-α [5] extends DiTs [30] for text-conditioned image generation. More recently, SD3 [11] propose MM-DiT design and find it scales well.

3. Scaling Denoising Backbone

3.1. Existing UNet Design

The UNet in diffusion models adopts a stack of residual blocks and a sequence of downsampling and upsampling convolutions, along with additional spatial attention layers at multiple resolutions [10, 16]. Recent T2I frameworks [9, 31, 34] mostly employ the ideas in simple diffusion [18] to improve the efficiency of UNet, i.e., tweaking more parameters and computation at smaller resolutions. Fig. 2 gives a comparison of the UNets for SD2 and SDXL. SDXL improves over SD2 in multiple dimensions: a) **Less downsampling rates.** SD2 uses [1, 2, 4, 4] as the multiplication rates to increase channels at different downsampling levels. DeepFloyd adopts [1, 2, 3, 4] to reduce computation, while SDXL uses [1, 2, 4], which completely removes the 4th downsampling level. b) **Cross-attention only at lower resolution.** Cross-attention is only computed at certain downsampling rates, e.g., SD2 applies cross-attention at first three downsampling rates ($1\times$, $2\times$, $4\times$), while SDXL only integrates text embedding at the $2\times$ and $4\times$ downsampling levels. c) **More compute at lower resolution.** SDXL applies more transformer blocks at the $2\times$ and $4\times$ downsampling levels, while SD2 applies uniform single transformer block at all three downsampling levels.

3.2. Controlled Comparison of UNets

To allow fair comparison of different UNets, we train all backbone variants in the same controlled settings, including the same *dataset*, *latent space*, *text-encoder* and *training settings*. Below we introduce the training configurations and evaluation metrics, based on which we compare all different backbones variants.

Training We train models on our curated dataset *LensArt*, which contains 250M text-image pairs (details in Sec 4). We use SDXL's VAE and the OpenCLIP-H [20] text encoder (1024 dim), without adding extra embedding layer or other conditioning. We train all models at 256×256 resolution with batch size 2048 upto 600K steps. We follow the setup of LDM [34] for DDPM schedules. We use AdamW [27] optimizer with 10K steps warmup and then constant learning rate 8e-5. We employ mixed precision training with BF16 and enables FSDP for large models.

Inference and Evaluation We use DDIM sampler [37] in 50 steps with fixed seed and CFG scale (7.5) for inference. To understand the training dynamics, we monitor the evolution of five metrics during training. We find the the metrics at early stage of training can help predict final model performance. Specifically, we measure composition ability and image quality with metrics including: 1) **TIFA** [19], which measures the faithfulness of a generated image to its text input via visual question answering (VQA). It contains 4K collected prompts and corresponding question-answer pairs generated by a language model. Image faithfulness is calculated by checking whether existing VQA models can answer these questions using the generated image. TIFA allows for fine-grained and interpretable evaluations of generated images. 2) **ImageReward** [40] which was learned to approximates human preference. We calculate the average ImageReward score over images generated with MSCOCO-10K prompts. Though ImageReward is not a normalized score, its scores are in the range of [-2, 2] and the average score over the number of images gives meaningful statistics to allow comparision across models. Due to space

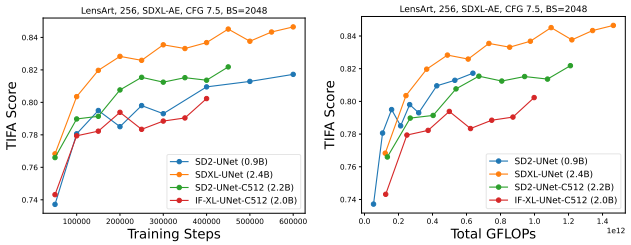

Figure 3. The evolution of TIFA score during training with different UNets on the same dataset in terms of training steps and training compute (GFLOPs). The compute FLOPs is estimated with 3× FLOPs of single DDPM step × batch size × steps.

constraints, we mainly show TIFA and ImageReward and provide results of other metrics (CLIP score [14, 32], FID, HPSv2 [39]) in Appendix.

SDXL vs SD2 vs IF-XL We compare the design of several existing UNet models from SDXL [31], DeepFloyd-IF [9], SD2 [34] and its scaled version in the above controlled settings. Specifically, we compare a) SD2 UNet (0.9B) b) SD2 UNet with 512 initial channels (2.2B) c) SDXL's UNet (2.4B) d) DeepFloyd's IF-XL UNet with 512 channels (2.0B). Fig. 3 shows that the naively scaled SD2-UNet (C512, 2.2B) achieves better TIFA score than the base SD2 model at the same training steps. However, the convergence speed is slower in terms of training FLOPs, which indicates *increasing channels is an effective but not an efficient approach*. SDXL's UNet achieves 0.82 TIFA within 150K steps, which is 6× faster than SD2 UNet and 3× faster than SD2-C512 in training iterations. Though its training iteration speed (FLOPS) is 2× slower than SD2, it still achieves the same TIFA score at 2× reduced training cost. SDXL UNet also can get a much higher TIFA score (0.84) with a clear margin over other models. Therefore SDXL's UNet design is *significantly better than others in terms of performance and training efficiency, pushing the Pareto frontier*.

3.3. Ablation of UNet Design

Now we have verified SDXL has a much better UNet design than SD2 and DeepFloyd variants. The question is why it excels and how to further improve it effectively and efficiently. Here we investigate how to improve SDXL's UNet by exploring its design space.

Search Space Table 1 shows different UNet configurations, and their corresponding compute complexity at 256 resolution. We mainly vary the initial channels and transformer depth. To understand the impact of each dimension of the design space, we select a subset of the variant models and train them with the same configurations. This forms our main "search space" for the UNet architecture. More ablations on the impact of VAE, training iterations and batch size can be found in Appendix.

The Effect of Initial Channels We train the following SDXL UNet variants with different channel numbers: 128, 192, and 384, with parameters 0.4B, 0.9B and 3.4B, respectively. Fig. 4 (a) shows that UNet with reduced channels from 320 to 128 still can outperform SD2's UNet with 320 channels, which shows that *less channel number can achieve better quality with proper architecture design*. However, the TIFA (also ImageReward/CLIP) scores are worse in comparison with SDXL's UNet, which indicates its importance in visual quality. Increasing channel number from 320 to 384 boosts the number of parameters from 2.4B to 3.4B. It also achieves better metrics than baseline 320 channels at 600K training steps. Note that the initial channel number C actually connects with other hyperparameters of the UNet, e.g., 1) the dimension of timestep embedding T is $4C$; 2) the number of attention head is linear with channel numbers, i.e. $C/4$. As shown in Table 1, the proportion of compute for attention layers are stable (64%) when C changes. This explains why increasing the width of UNet

Table 1. Comparing UNet variants in terms of their hyperparameter, number of parameters, and inference complexity (GMACs). We also list the portion of compute allocated for attention operations. The original architecture hyperparameters are marked in **bold**.

UNet	Channels	Channel Mult.	Res. Blocks	Atten. Res.	Tran. Depth	Params (B)	GMACs		
							Total	Atten.	Atten.%
SD2 [34]	**320**	**[1,2,4,4]**	**2**	**[4, 2, 1]**	**[1, 1, 1]**	**0.87**	**86**	34	**39**
	512					2.19	219	85	39
IF-XL [9]	512	**[1,2,3,4]**	**3**	**[4, 2, 1]**	**[1, 1, 1]**	2.04	194	23	12
	704					3.83	364	42	**12**
SDXL [31]	128	**[1, 2, 4]**	**2**	**[4, 2]**	**[0, 2, 10]**	0.42	35	23	65
	192					0.90	75	48	65
	320					**2.39**	**198**	127	**64**
	384					3.40	282	179	64
	320	**[1, 2, 4]**	**2**	**[4, 2]**	[0, 2, 2]	0.85	98	43	44
					[0, 2, 4]	1.24	123	64	52
					[0, 2, 12]	2.78	223	147	66
					[0, 2, 14]	3.16	248	168	68
					[0, 4, 4]	1.32	143	84	59
					[0, 4, 8]	2.09	193	123	64
					[0, 4, 12]	2.86	243	167	69
	384	**[1, 2, 4]**	**2**	**[4, 2]**	[0, 4, 12]	4.07	346	237	69

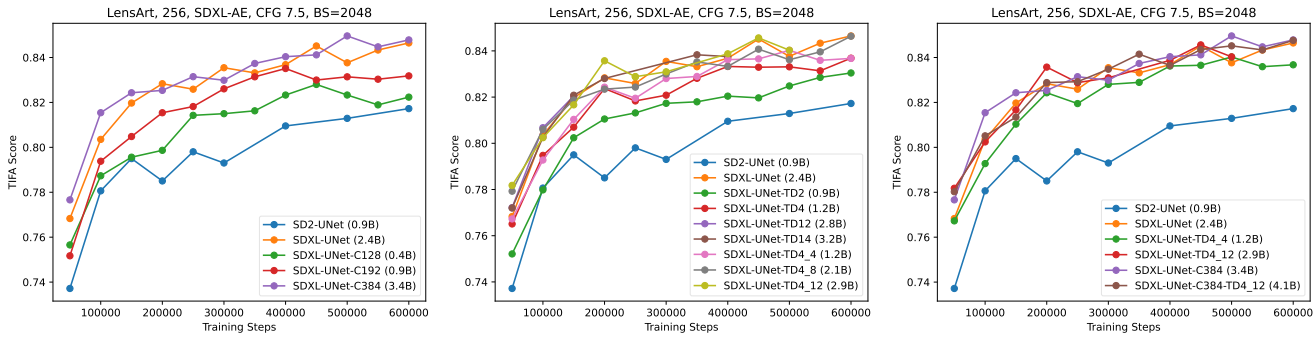

Figure 4. Evolution of TIFA score during training with scaled UNet variations. The baseline models are UNets of SD2 and SDXL. We train SDXL UNet variants with changes in (a) channels C (b) transformer depth (TD) 3) both channels and TD.

also brings alignment improvement as shown in Fig. 4.

The Effect of Transformer Depth The transformer depth (TD) setting controls the number of transformer blocks at certain input resolution. SDXL applylies 2 and 10 transformer blocks at $2\times$ and $4\times$ downsampling level, respectively. To understand its effect, we trained the variants shown in Table 1 with different TDs, ranging from 0.9B to 3.2B parameters. Specifically, we first change the TD at the $4\times$ downsampling rate, obtaining TD2, TD4, TD12 and TD14, then we further change the depth at $2\times$ downsmapling rate, resulting in TD4_4, TD4_8 and TD4_12. Note the portion of attention operations also increases with TDs accordingly. Fig. 4 (b) shows that increasing TD at $4\times$ downsampling rate from 2 to 14 continuously improves TIFA score. From the comparison between TD4 and TD4_4, we see that increasing transformer depth at $2\times$ resolution ($2 \rightarrow 4$) also improves TIFA score. TD4_4 has competitive performance in comparison with SDXL's UNet while having 45% less parameters and 28% less compute for inference. In Appendix, we show TD4_4 achieves same TIFA score $1.7\times$ faster than SDXL UNet in terms of wall-clock training time. TD4_8 has almost the same performance as SDXL's UNet with 13% less parameters. Since the text-image alignment (TIFA) is mostly about large objects in the image, it is helpful to *allocate more cross compute at lower resolution or global image level beyond efficiency considerations*.

Scaling both Channels and Transformer Depth Given the effect of channels and transformer depth, we further explored enlarging both the channel numbers (from 320 to 384) and transformer depth ([0, 2, 10] $\rightarrow$ [0, 4, 12]). Fig. 4 (c) shows that it achieves slightly higher TIFA scores during training than SDXL-UNet. However, the advantage over simply increasing channels or transformer depth is not apparent, which means *there is a performance limit for models to continue scale under metrics like TIFA*.

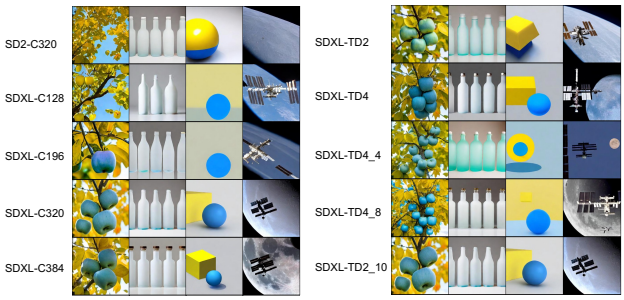

Figure 5. Visualizing the effect of UNet scaling on text-image alignment. We change the UNet along two dimensions: channel number (left) and transformer depth (right). The prompts are: 1) "square blue apples on a tree with circular yellow leaves" 2) "five frosted glass bottles" 3) "a yellow box to the right of a blue sphere" 4) "the International Space Station flying in front of the moon"

Visualizing the Effect of UNet Scaling Fig. 5 shows the images generated by different UNets with the same prompts. We can see that as the channel number or transformer depth increases, the images become more aligned with the given prompts (e.g., color, counting, spatial, object). Some images generated by certain UNet variant are better than the original SDXL UNet (C320), i.e., SDXL-C384 and SDXL-TD4_8 both generate more accurate images with the 4th prompt.

3.4. Comparing with Transformers

DiT [30] demonstrates that scaling up the transformer complexity can get consistently improved image fidelity for class-conditioned image generation on ImageNet. PixArt-α [5] extends DiT to text-conditioned image generation with similar backbone. However, there is a lack of fair comparison with UNet in a controlled setting. To compare with UNet and understand its scaling property, we train multiple scaled version of PixArt-α, keeping other components and settings the same as previous ablations. Table 2 shows the configuration of our scaled variants. The difference with the original PixArt-α model lies at: 1) we use SDXL's VAE instead of

Table 2. Hyperparameter settings for Transformer-based backbones at resolution 256x256. The original PixArt-α [5] model uses T5-XXL tokenizer, while we use OpenCLIP-H to keep consistency with UNet experiments. The original architecture settings are marked in **bold**. p, h and d denote patch size, hidden dimension, and depth.

Model	VAE	p	h	d	#heads	Text Encoder	Max Tokens	Token Dim.	Cap. Emb	GMACs	Params (B)
PixArt-α-XL/2[5]	**SD2**	**2**	**1152**	**28**	**16**	**4.3B Flan-T5-XXL**	**120**	**4096**	**Y**	139	0.61
Ours	SDXL	**2**	**1152**	**28**	**16**	354M OpenCLIP-H	77	1024	**Y**	139	0.61
			1536						**Y**	247	1.08
			1024	**28**	**16**				N	110	0.48
				56					N	220	0.95

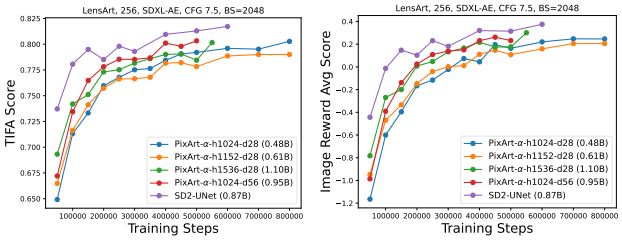

Figure 6. The evolution of TIFA and ImageReward during training for scaled PixArt-α variants as well as SD2 UNet.

SD2's VAE; 2) we use OpenCLIP-H text encoder instead of T5-XXL [7], the token embedding dimension is reduced from 4096 to 1024. The token length is 77 instead of 120.

Ablation Space We ablate the PixArt-α model in the following dimensions: 1) *hidden dimension* h: PixArt-α inherits the design of DiT-XL/2 [30], which has 1152 dimension. We also consider 1024 and 1536. 2) *transformer depth* d: we scale the transformer depth from 28 to 56. 3) *caption embedding*: the caption embedding layer maps the text encoder output to dimension h. When the hidden dimension is the same as text embedding (i.e., 1024), we can skip the caption embedding to use the token embedding directly.

The Effect of Model Scaling As shown in Fig. 6, scaling the hidden dimension h and model depth d both result in improved text-image alignment and image fidelity, while scaling depth d changes model's compute and size linearly. Both $d56$ and $h1536$ variants achieve $\tilde{1}.5\times$ faster convergence speed than the baseline $d28$ model with similar parameter size and compute.

Comparison with UNet The PixArt-α variants yield lower TIFA and ImageReward scores in comparison with SD2-UNet trained in same steps, e.g., SD2 UNet reaches 0.80 TIFA and 0.2 ImageReward at 250K steps while the 0.9B PixArt-α variant gets 0.78 and 0.1. PixArt-α [5] also reports that training without ImageNet pre-training tends to generate distorted images in comparison to models initialized from pre-trained DiT weights, which is trained 7M steps on ImageNet [30]. Though DiT [30] proves the UNet is not a must for diffusion models, the PixArt-α variants take longer iterations and more compute to achieve similar performance as UNet. We leave this improvement for future works and expect architecture improvement can mitigate this issue, such as works done in [11, 12, 41].

4. Scaling Training Data

4.1. Dataset Curation

We curate our own datasets named *LensArt* and *SSTK*. *LensArt* is a 250M image-text pairs sourced from 1B noisy web image-text pairs. We apply a series of automatic filters to remove data noise, including but not limited to NSFW content, low aesthetic images, duplicated images, and small images. *SSTK* is another internal dataset with about 350M cleaned data. Table 3 shows the statistics of the datasets. More detailed analysis can be seen in the Appendix.

Table 3. Dataset statistics. **I-C** indicates the number of unique image-caption pairs. **AE** indicates the average aesthetic score for the dataset. **I-N** indicates the total number of image-noun pairs, where each pair includes the image and a noun that is unique within corresponding real and synthetic captions. **UN** indicates the number of unique nouns within whole text corpus (here nouns includes both nouns and proper nouns for simplicity). **N/I** indicates the average number of nouns per image. **w. Syn** indicates the dataset incorporates synthetic captions.

Datasets	**I**	**AE**	**I-N**	**UN**	**N/I**
LensArt-raw	1.0B	5.00	7.1B	3.9M	7.1
LensArt	250M	5.33	1.8B	1.2M	7.0
SSTK	360M	5.20	2.2B	680K	6.0
LensArt + SSTK	610M	5.25	3.9B	1.7M	6.5
LensArt (w. Syn)	250M	5.33	3.2B	1.3M	12.8
SSTK (w. Syn)	360M	5.20	4.2B	1.1M	11.6
LensArt + SSTK (w. Syn)	610M	5.25	7.3B	2.0M	12.2

4.2. Data Cleaning

The quality of the training data is the prerequisite of data scaling. Compared to training with noisy data source, a high-quality subset not only improves image generation quality, but also preserves the image-text alignment. LensArt is $4\times$ smaller than its unfiltered 1B data source, with hundreds of million noisy data removed. However, model trained with this high-quality subset improves the average aesthetic score [23] on the generated images from 5.07 to 5.20. This is caused by LensArt having an average aesthetic score of

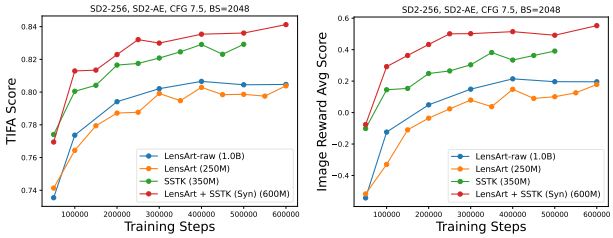

Figure 7. SD2 models trained on different datasets and their corresponding TIFA and average ImageReward scores during training. Increasing the scale of the dataset by combining LensArt and SSTK gets the best results.

Alt-text
- English muffins

Top-5 synthetic captions
- A white basket filled with english muffins on top of a table.
- A white basket filled with english muffins is placed on top of a table.
- A white basket filled with english muffins on top of a red napkin.
- A white basket filled with english muffins is on top of a table.
- A white basket filled with english muffins on top of a red cloth.

Figure 9. Synthetic captions provide descriptions with more details.

5.33, higher than 5.00 in LensArt-raw. Moreover, as shown in Fig. 7, a SD2 model trained in LensArt achieves similar TIFA score in comparison the one trained with the raw version, demonstrating that the filtering does not hurt image-text alignment. The reason is that sufficient commonsense knowledge is still retained under aggressive filtering while enormous duplicated and long-tail data removed.

4.3. Expanding Knowledge via Synthetic Captions

To increase the valid text supervision for the smaller yet higher-quality data, we adopt an internal image captioning model, similar to BLIP2 [24], to generate synthetic captions. The captioning model produces five generic descriptions for each image ranked by prediction confidence as in Fig. 9. One of the 5 synthetic captions and original alt-text is randomly selected to pair with the image for model training under 50% chance. Thus we double the image-text pairs and significantly increase the image-noun pairs as shown in Table 3. Thanks to the text supervision expanding by synthetic captions, the image-text alignment and fidelity can be consistently boosted, as shown in Table 4. Specifically, the ablation on LensArt shows that synthetic captions significantly improves ImageReward score. In addition, we find that randomly selecting one in top-5 synthetic captions is slightly

Table 4. Synthetic caption ablations by training SD2 for 250K steps. **IR**: ImageReward. **Top5 Syn.**: randomly select one in top-5 synthetic captions ranked by the caption prediction confidences; **Top1 Syn.**: only select top-1 synthetic caption

Model	Synthetic caption	TIFA	CLIP	IR	FID
SD2	LensArt	0.810	0.269	0.345	**17.9**
	LensArt + Top5 Syn.	**0.835**	0.270	**0.524**	18.9
	LensArt + Top1 Syn.	0.833	**0.271**	0.513	18.3

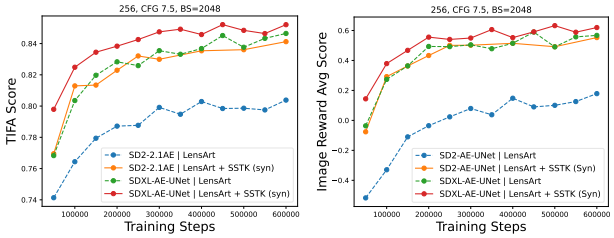

Figure 8. Comparing SD2 and SDXL trained on LensArt and LensArt+SSTK (w.Syn). Enlarging training set helps improving model performance. Stronger model yields better performance with larger dataset.

better than always selecting top-1, which is adopted as the default training scheme for synthetic captions. Different with PixArt-α which always replaces the original captions with long synthetic captions, we provide an alternate way to scale up the image-text pairs by random flipping captions, which is in concurrent with the caption enhancement work of DALL-E 3 [3].

4.4. Data Scaling Increases Training Efficiency

Combined datasets The text-image alignment and image quality can be further improved as the dataset scale increases. Here we compare the SD2 models trained on different datasets and compare their convergence speed: 1) LensArt 2) SSTK and 3) LensArt + SSTK with synthetic caption. We also compare training with unfiltered LensArt-raw as the baseline. Fig. 7 shows that *combining LensArt and SSTK significantly improves the convergence speed and the upper limit of two metrics in comparison with models trained on LensArt or SSTK only*. SDXL model trained on LensArt + SSTK reaches 0.82+ TIFA score in 100K steps, which is 2.5$\times$ faster than SDXL trained with only LensArt.

Advanced models scale better on larger dataset Fig. 8 shows that SD2 model can get significant performance gain when training on the scaled (combined) dataset. SDXL still gets performance gain over the SD2 model when trained with the scaled dataset, indicating that models with large capacity have better performance when the dataset scale increases.

5. More Scaling Properties

Relationship between performance and model FLOPs Fig. 11 (a-b) shows the correlation between TIFA score obtained at fixed steps (i.e., 600K) and model compute complexity (GFLOPs) as well as model size (#Params) for all examined SD2 and SDXL variants. We see the TIFA score correlates slightly better with FLOPs than parameters, indicating the importance of model compute when training budget is sufficient, which aligns with our findings in Sec 3.

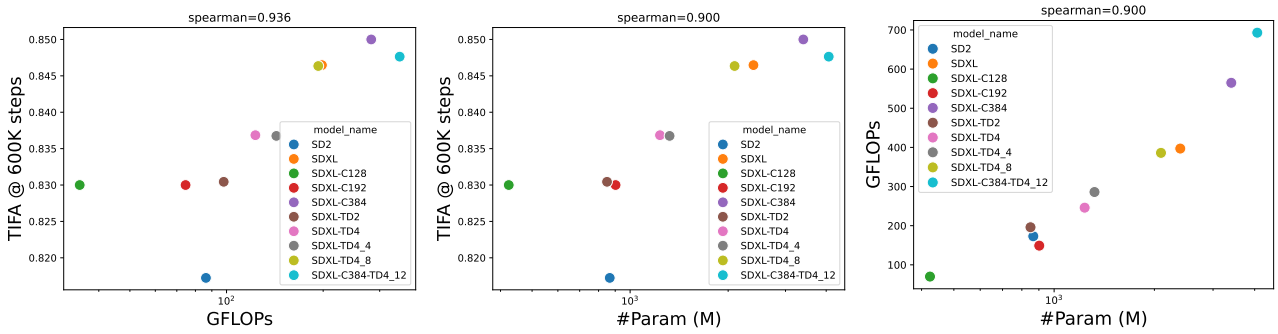

Figure 10. (a-b) The correlation of TIFA score with UNet's inference complexity (GFLOPs) and number of parameters when trained with fixed steps. (c) shows the correlation between model parameters and FLOPs.

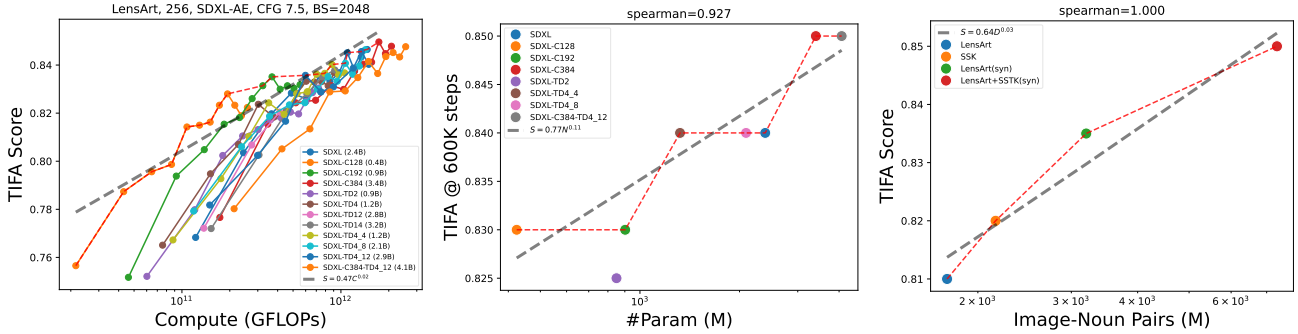

Figure 11. Fitting the scaling law of TIFA performance S as a function of the training compute C and model size N, and dataset size D, with the training history of SDXL variants and SD2 in fixed steps. The (- -) indicates the Pareto frontier of the scaling graph.

Relationship between performance and data size Fig. 11(c) shows the correlation between SD2's TIFA score and dataset size in terms of number of image-noun pairs. Each image-noun pair is defined as an image paired with one noun in its caption. It measures the interaction between the fine-grained text unit with the image. We see a linear correlation between TIFA and the scale of image-noun pairs when scaling up the cleaned data. Compared to LensArt-raw with similar amount of image-noun pairs, LensArt+SSTK is much better, which indicates the importance of data quality.

Numerical Scaling Law The scaling law of LLMs [17, 21] reveals that LLM's performance has a precise power-law scaling as a function of dataset size, model size, and compute budget. Here we fit similar scaling functions for SDXL variants and SD2. The TIFA score S can be a function of total compute C (GFLOPs), model parameter size N (M parameters) and dataset size D (M image-noun pairs) as shown in Fig. 11. Specifically, with the Pareto frontier data points, we can fit the power-law functions to be $S = 0.47C^{0.02}$, $S = 0.77N^{0.11}$, and $S = 0.64D^{0.03}$, which approximate the performance in a range given sufficient training. Similar as LLMs, we see larger models are more sample efficient and smaller models are more compute efficient.

Model Evaluation at Low Resolution One may wonder whether the models' relative performance will change at high resolution training, so that the gap between models can be mitigated. In the Appendix, we show continue training models at 512 resolution slightly improve their 256 resolution metrics without significant change. Though image quality and aesthetics can be improved via high-quality fine-tuning [8], it is hard for the inferior model to surpass when trained on the same data, especially when the high resolution data is much less than its lower resolution version. The majority composition capability is developed at low resolution, which enables us to assess model's performance at the early stage of low resolution training.

6. Conclusions

We present a systematic study on the scaling properties of training diffusion based T2I models, including the effects of scaling both denoising backbone and dataset. Our study demonstrates practical paths to improve T2I model performance by properly scaling up existing denoising backbones with large-scale datasets, which results in better text-image alignment and image quality, as well as training efficiency. We hope those findings can benefit the community for pursuing more scaling-efficient models.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1

[2] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *CVPR*, 2023. 2

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. *https://cdn.openai.com/papers/dall-e-3.pdf*, 2023. 2, 7

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 2, 3, 5, 6

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1

[7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 6

[8] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 2, 8

[9] DeepFloyd. Deepfloyd. *https://github.com/deep-floyd/IF*, 2023. 1, 2, 3, 4

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2, 3

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 3, 6

[12] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. 2023. 2, 6

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4, 11

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[17] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 1, 8

[18] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 3

[19] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 3, 11

[20] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 3

[21] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1, 8

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[23] LAION. Laion aesthetic v2. *https://github.com/christophschuhmann/improved-aesthetic-predictor*, 2022. 6

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 7

[25] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *arXiv preprint arXiv:2306.00980*, 2023. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 11

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 3

[28] Alexander Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[29] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2, 3, 5, 6

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 3, 4, 12

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 4, 11

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3, 2022. 1, 2

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2021. 1, 2, 3, 4

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 2

[36] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 11

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[39] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 4, 11

[40] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023. 3, 11

[41] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. *arXiv preprint arXiv:2311.18257*, 2023. 6

[42] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 2

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

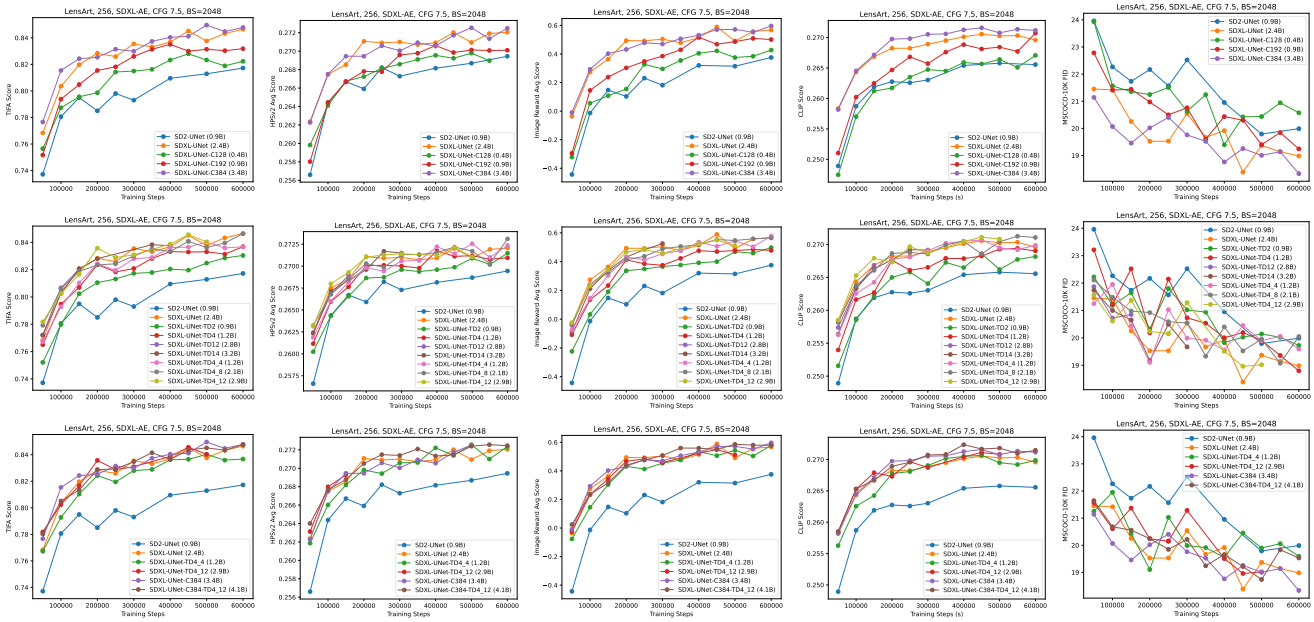

Figure 12. The evolution of all metrics during training for UNet variants. The baseline models are the UNets of SD2 and SDXL. All models are trained with SDXL VAE at 256×256 resolution. The 1st row shows SDXL UNets with different initial channels. The 2nd row shows SDXL UNets with different TDs. The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

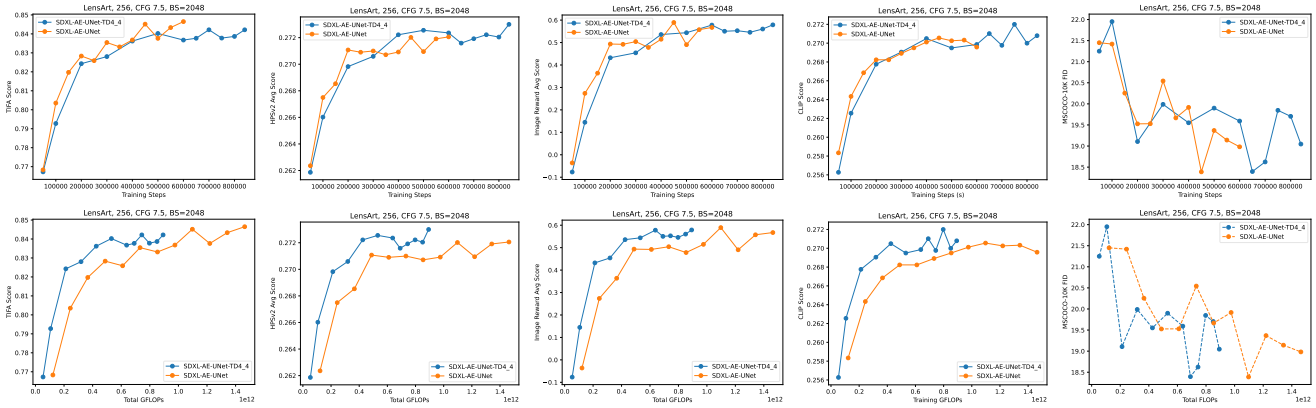

Figure 13. Comparing metrics evolution speed of SDXL UNet and its TD4_4 variant in terms of training steps and total compute (GFLOPs). TD4_4 achieves similar or better metric scores at much less training cost.

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.

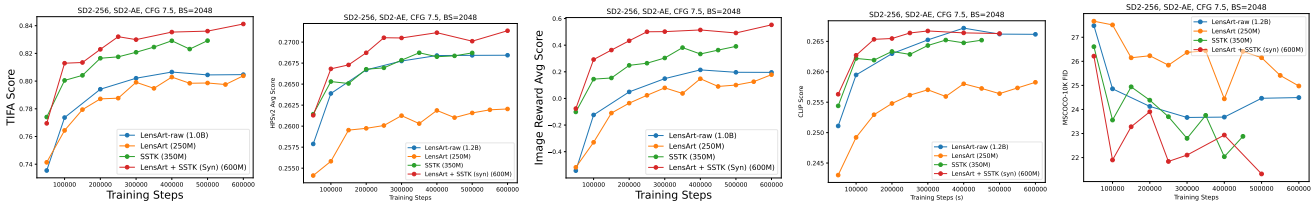

Figure 14. Training SD2 model with different datasets. All metrics show that LensArt + SSTK has better scores than LensArt or SSTK only. Note that the HPSv2 and CLIP scores for LensArt-Raw are much higher than LensArt. The reason is that unfiltered dataset tends to generate images with more meme text.

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.

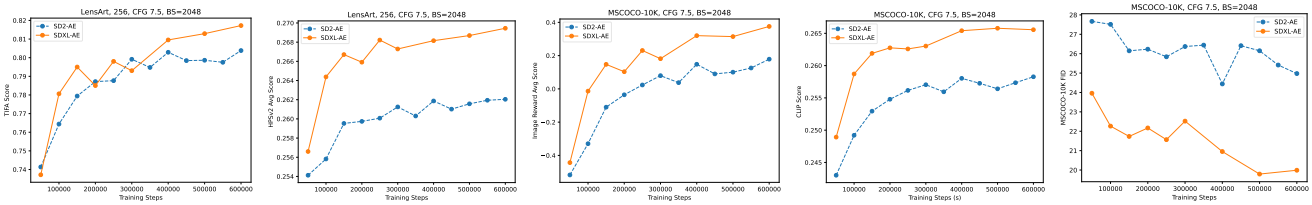

Figure 15. Training SD2 UNet model with different VAEs. The SDXL's VAE has significant improvement on all metrics over SD2's VAE.

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.

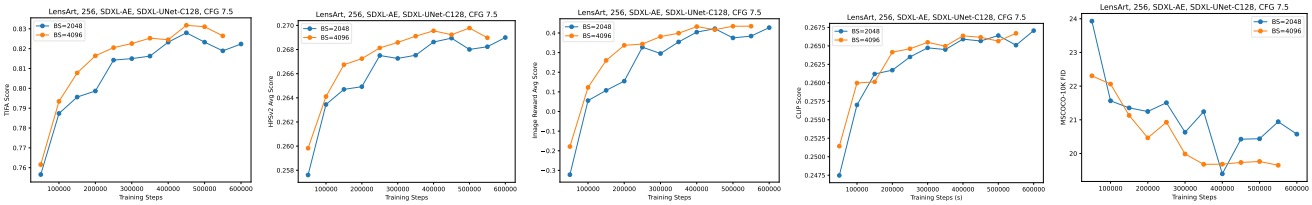

Figure 16. Training SDXL-UNet-C128 with different batch sizes.

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.

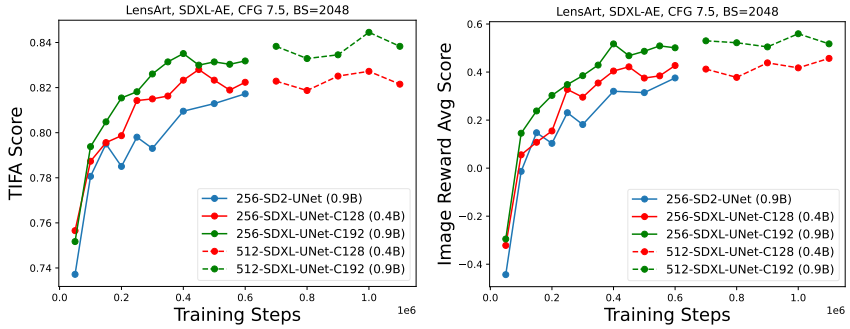

Figure 17. TIFA and ImageReward do not change much during high resolution fine-tuning stage (dashed lines)

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.

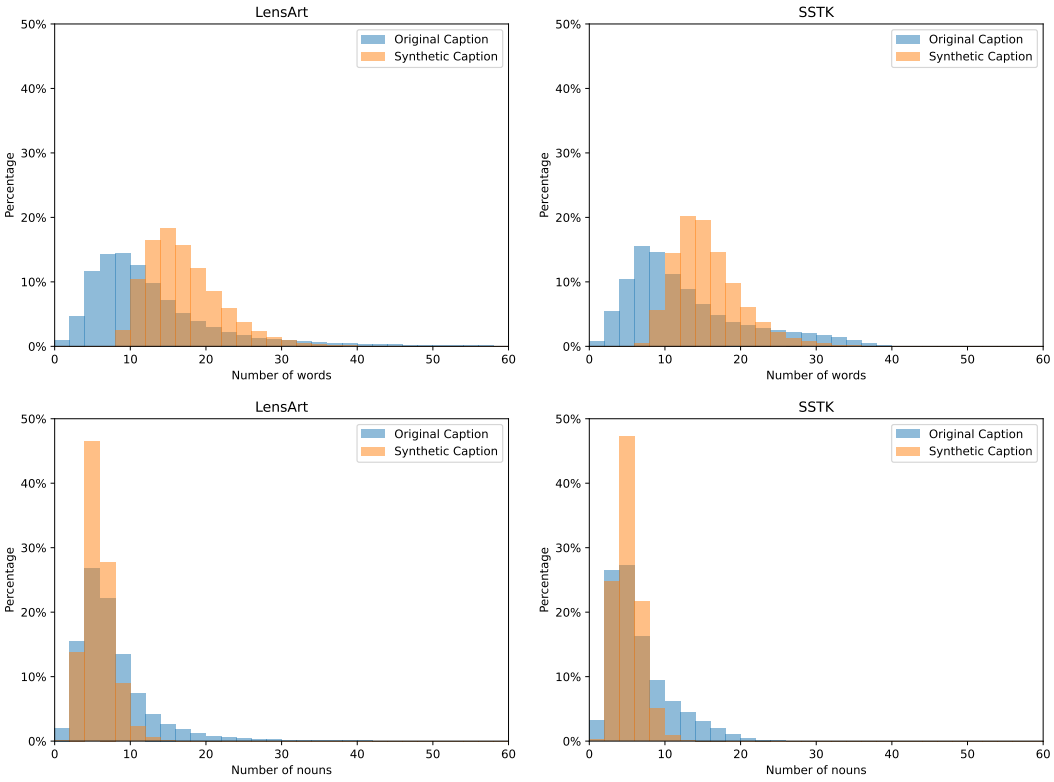

Figure 18. Histograms of word and noun numbers in the original and synthetic captions of different datasets