DATA EXTRAPOLATION FOR TEXT-TO-IMAGE GENERATION ON SMALL DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image generation requires large amount of training data to synthesizing high-quality images. For augmenting training data, previous methods rely on data interpolations like cropping, flipping, and mixing up, which fail to introduce new information and yield only marginal improvements. In this paper, we propose a new data augmentation method for text-to-image generation using linear extrapolation. Specifically, we apply linear extrapolation only on text feature, and new image data are retrieved from the internet by search engines. For the reliability of new text-image pairs, we design two outlier detectors to purify retrieved images. Based on extrapolation, we construct training samples dozens of times larger than the original dataset, resulting in a significant improvement in text-to-image performance. Moreover, we propose a NULL-guidance to refine score estimation, and apply recurrent affine transformation to fuse text information. Our model achieves FID scores of 7.91, 9.52 and 5.00 on the CUB, Oxford and COCO datasets. The code and data will be available on GitHub.

1 INTRODUCTION

Text-to-image generation aims to synthesize images according to textual descriptions. As the bridge between human language and generative models, text-to-image generation (Reed et al., 2016b; Ye et al., 2023; Sauer et al., 2023; Rombach et al., 2022; Ramesh et al., 2022)is applied to more and more application domains, such as digital human (Yin & Li, 2023), image editing (Brack et al., 2024), and computer-aided design (Liu et al., 2023). The diversity of applications leads to a large number of small datasets, where existing data are not sufficient to train high-quality generative models, and generative large models cannot overcome the long-tail effect of diverse applications.

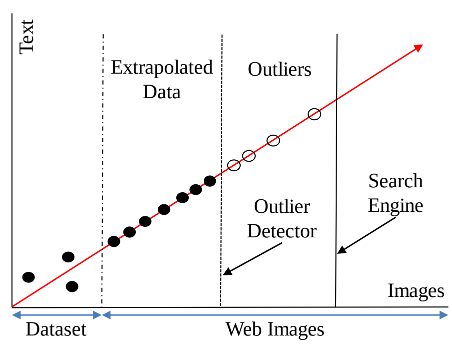

Figure 1: An illustration of data linear extrapolation. We use search engine and outlier detectors to ensure the image similarity. Extrapolation produces much more text-image pairs than the original dataset.

To augment training data, existing methods typically rely on data interpolation techniques such as cropping, flipping, and mixing up images (Zhang et al., 2017). While these methods leverage human knowledge to create new perspectives on existing images or features, they do not introduce new information and yield only marginal improvements. Additionally, Retrieval-base models (Chen et al., 2022; Sheynin et al., 2022; Li et al., 2022) employs retrieval methods to gather relevant training data from external databases like WikiImages. However, these external databases often contain very few images for specific entries, and their description styles differ significantly from those in text-to-image datasets. Furthermore, VQ-diffusion (Gu et al., 2022) pre-trains its text-to-image model on the Conceptual Caption dataset with 15 million images, but the resulting improvements are not obvious.

In this paper, we explore data linear extrapolation to augment training data. Linear extrapolation can be risky, as similar text-image pairs may not be nearby in Euclidean space. For information reliability, as depicted in Figure 1, we explore linear extrapolation only on text data, and new image data are retrieved from the internet by search engines. And then outlier detectors are designed to purify retrieved web images. In this way, the reliability of new text-image pairs are guaranteed by search precision and outlier detection.

To detect outliers from web images, we divide outliers into irrelevant and similar ones. For detecting irrelevant outliers, K-means (Lloyd, 1982) algorithm is used to cluster noisy web images into similar images and outliers. In the image feature space generated by a CLIP encoder (Radford et al., 2021), similar images will be close to dataset images, while outliers will be far away. Based on this observation, we remove images that differ significantly from dataset images. For detecting similar outliers, each web image is assigned a label by a fine-grained classifier trained on the original dataset. If the label does not match the search keyword, the image is considered as an outlier and removed. For every purified web image, we extrapolate a new text descriptions according to the local manifold of dataset images. Based on extrapolation, we construct training samples dozens of times larger than the original dataset.

Moreover, we propose NULL-condition guidance to refine score estimation for text-to-image generation. Classifier-free guidance (Ho & Salimans, 2022) uses a dummy label to refine label-conditioned image synthesis. Similarly, in text-to-image generation, such a dummy label can be replaced by a prompt with no new physical meaning. For example, “a picture of bird” provides no information for the CUB dataset “a picture of flower” provides no information for the Oxford dataset). In addition, we apply recurrent affine transformation (RAT) in the diffusion model for handling complex textual information.

The contributions of this paper are summarized as follows:

- We propose a new data augmentation method for text-to-image generation using linear extrapolation. Specifically, we apply linear extrapolation only on text feature, and new image data are retrieved from the internet by search engines.
- We propose a NULL-condition guidance to refine the score estimation for text-to-image generation. This guidance is also applicable to existing text-to-image models without further training.
- We apply recurrent affine transformation in the diffusion model for handling complex textual information.

2 Related work

GAN-based text-to-image models. Text-to-image synthesis is a key task within conditional image synthesis (Feng et al., 2022; Tan et al., 2022; Peng et al., 2021; Hou et al., 2022). The pioneering work of (Reed et al., 2016b) first tackled this task using conditional GANs (Mirza & Osindero, 2014). To better integrate text information into the synthesis process, DF-GAN (Tao et al., 2022) introduced a deep fusion method featuring multiple affine layers within a single block. Unlike previous approaches, DF-GAN eliminated the normalization operation without sacrificing performance, thus reducing computational demands and alleviating limitations associated with large batch sizes. Building on DF-GAN, RAT-GAN employed a recurrent neural network to progressively incorporate text information into the synthesized images. GALIP (Tao et al., 2023) and StyleGAN-T (Sauer et al., 2023) explore the potential of combining GAN models with transformers for large-scale text-to-image synthesis. However, the aforementioned GAN-based models often struggle to produce high-quality images.

Diffusion-based text-to-image models. Recently, diffusion models (Ho et al., 2020; Song & Ermon, 2019; Song et al., 2021; Hyvärinen, 2005) have demonstrated impressive generation performance across various tasks. Building on this success, Imagen (Saharia et al., 2022) and DALL·E 2 (Ramesh et al., 2022) can synthesize images that are sufficiently realistic for real-world applications. To alleviate computational burdens, they first generate 64×64 images and then upsample them to high-resolution using another diffusion model. Additionally, the Latent Diffusion Model (Rombach et al., 2022) encodes high-resolution images into low-resolution latent codes, avoiding the exponential computation costs associated with increased resolution. DiT (Peebles & Xie, 2023) integrated latent diffusion models and transformers to enhance performance on large datasets. VQ-Diffusion (Gu et al., 2022) pre-train their text-to-image model on the Conceptual Caption dataset, which contains 15 million text-image pairs, and then fine-tune it on smaller datasets like CUB, Oxford, and COCO. Hence, VQ-Diffusion is the work most similar to ours but we use significantly less pre-training data while achieving better results.

Data augmentation methods. Data augmentation increases training data to improve the performance of deep learning applications, from image classification (Krizhevsky et al., 2012) to speech recognition (Graves et al., 2013; Amodei et al., 2016). Common techniques include rotation, translation, cropping, resizing, flipping (LeCun et al., 2015; Vedaldi & Zisserman, 2016), and random erasing (Zhong et al., 2020) to promote visually plausible invariances. Similarly, label smoothing is widely used to boost the robustness and accuracy of trained models (Müller et al., 2019; Lukasik et al., 2020). Mixup (Zhang et al., 2017) involves training a neural network on convex combinations of examples and their labels. However, interpolated samples fail to introduce new information and effectively address data scarcity. Hence, Re-imagen (Chen et al., 2022; Sheynin et al., 2022; Li et al., 2022) retrieval relevant training data from external databases to augment training data.

3 Linear Extrapolation for Text-to-image Generation

In this section, we begin by collecting similar images from the internet. Next, we explain how to extrapolate text descriptions. Following that, we use the extrapolated text-image pairs to train a diffusion model with RAT blocks. Finally, we sample images using NULL-condition guidance.

3.1 Collecting Similar and Clean Images

Linear extrapolation requires the images to be sufficiently close in semantic space. Hence, we automatically retrieve similar images by searching for their classification labels. However, search engines return both similar images and outliers. To eliminate unwanted outliers, we employ a cluster detector for irrelevant outliers and a classification detector for similar outliers. For the cluster detector, each image is encoded into a vector using the CLIP image encoder. Images retrieved with the same keyword are then clustered using K-means. If the distance from the cluster center to dataset images exceeds a threshold, this cluster is excluded. For the classification detector, we train a fine-grained classification model on the original dataset, which assigns a label to each web image. If the label does not match with the search keyword, corresponding image is then excluded.

3.2 Linear Extrapolation on Text Feature Space

Here we introduce how to extrapolates text descriptions for web images. Assuming that web images are sufficiently close to dataset images in semantic space, each web image can be represented by nearest k images:

$$\arg\min_{W} |\mathbf{f} \ - \mathbf{F} \times \mathbf{w}|^2, \tag{1}$$

where $\mathbf{w} = [w_1, w_2, ..., w_k]$are the reconstruction weights and $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_k]$ are the image features of dataset images produced by CLIP image encoder. Since the above equation is a super-determined problem, we solve this coefficient using least squares:

$$\mathbf{w} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{f}. \tag{2}$$

We assume that the image feature space and text feature space share the same local manifold. Hence, the image reconstruction efficient **w** can be used to compute the text feature of web images:

$$\mathbf{s} = \mathbf{S} \times \mathbf{w}, \tag{3}$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_k]$ is the fake sentence features for nearest k dataset images, and **s** is the sentence feature for a web image.

3.3 Recurrent Diffusion Transformer on Latent Space

The training objective of the diffusion model is the squared error loss proposed by DDPM (Ho et al., 2020):

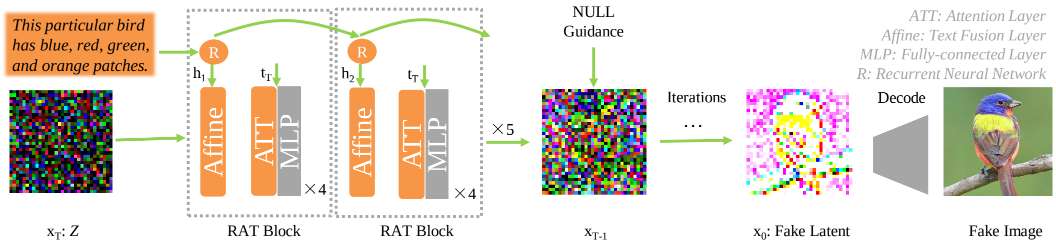

Figure 2: Latent diffusion model with recurrent affine transformation and NULL-guidance for text-to-image synthesis. The RAT blocks are connected by a recurrent neural network to ensure the global assignment of text information.

$$L(\theta) = \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\bar{\alpha}}_t} \mathrm{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \right) \right\|^2, \tag{4}$$

where $\epsilon \in N(0,1)$ is the score noise injected at every diffusion step, and ϵ_θ is the predicted noise by a diffusion network consisted of 12 transformer layers. $\bar{\bar{\alpha}}_t$ and $\bar{\alpha}_t$ are hyper-parameters controlling the speed of diffusion. The work of score mismatching (Ye & Liu, 2024) shows that predicting the score noise leads to an unbiased estimation.

Network architecture. As depicted in Fig 2, the diffusion network consists of transformer blocks. Recurrent affine transformation is used to enhance the consistency between transformer blocks. To avoid directly mixing text embedding and time embedding, we stack four transformer blocks as a RAT block and text embedding is fed into the top of each RAT block. Each RAT block applies a channel-wise shifting operation on a image feature map:

$$c' = c + \beta, \tag{5}$$

where c is the image feature vector and β is shifting parameters predicted by a one-hidden-layer multi-layer perception (MLP) conditioned on recurrent neural network hidden state h_t.

In each transformer block, we inject time embedding by a channel-wise scaling operation and a channel-wise shifting operation on c. At last, the image feature c is multiplied by a scaling parameter α. This process can be formally expressed as:

$$c' = Transformer((1+\gamma) \cdot c + \beta) \cdot \alpha, \tag{6}$$

where α, γ, β are parameters predicted by two one-hidden-layer MLPs conditioned on time embedding. When applied to an image feature map composed of $w \times h$ feature vectors, the same affine transformation is repeated for every feature vector.

Early stop of fine-tuning. Extrapolation may produces training data very close to the original dataset, which makes fine-tuning saturate very quickly. Excessive fine-tuning epochs would forget knowledge gained from the extrapolated data and overfit small datasets. As a result, the training loss of the diffusion model becomes unreliable. Therefore, fine-tuning should be stopped when the FID score begins to increase.

3.4 Synthesizing Fake Images

Finally, we introduces how to synthesizing images from scratch. As depicted in Figure 2, the synthesis begins with sampling a random vector z from standard Gaussian distribution. And then, this noise is gradually denoised into an image latent code by the diffusion model. The reverse diffusion iterations are formulated as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta \left(\mathbf{x}_t, t \right) \right) + \sigma_t \mathbf{z}, \tag{7}$$

where, α_t, $\bar{\alpha}_t$ and σ_t are diffusion hyper-parameters, and z is a random vector sampled from standard Gaussian distribution. At last, we decode image latent codes into images with the pre-trained decoder from Stable Diffusion (Rombach et al., 2022).

NULL guidance. A sentence with no new information is able to boost text-to-image performance obviously. This guidance is inspired by Classifier-free diffusion guidance (Ho & Salimans, 2022) which uses a dummy class label to boost label-to-image performance. Similarly, we design CLIP prompt without obvious visual meaning and embed them into the diffusion model. Specifically, we denote the original score estimation based on text description as ϵ_{text} and score estimation based on null description as ϵ_{null}. Then we mix these two estimations for a more accurate estimation ϵ':

$$\epsilon' = (\epsilon_{text} - \epsilon_{null}) \times \eta + \epsilon_{null}, \tag{8}$$

where, η is the guidance ration controlling the balance of two estimations. When $\eta = 1$, NULL Guidance falls back to an ordinary score estimation. Usually, a NULL prompt with the average meaning of the dataset achieve the best performance.

4 EXPERIMENTS

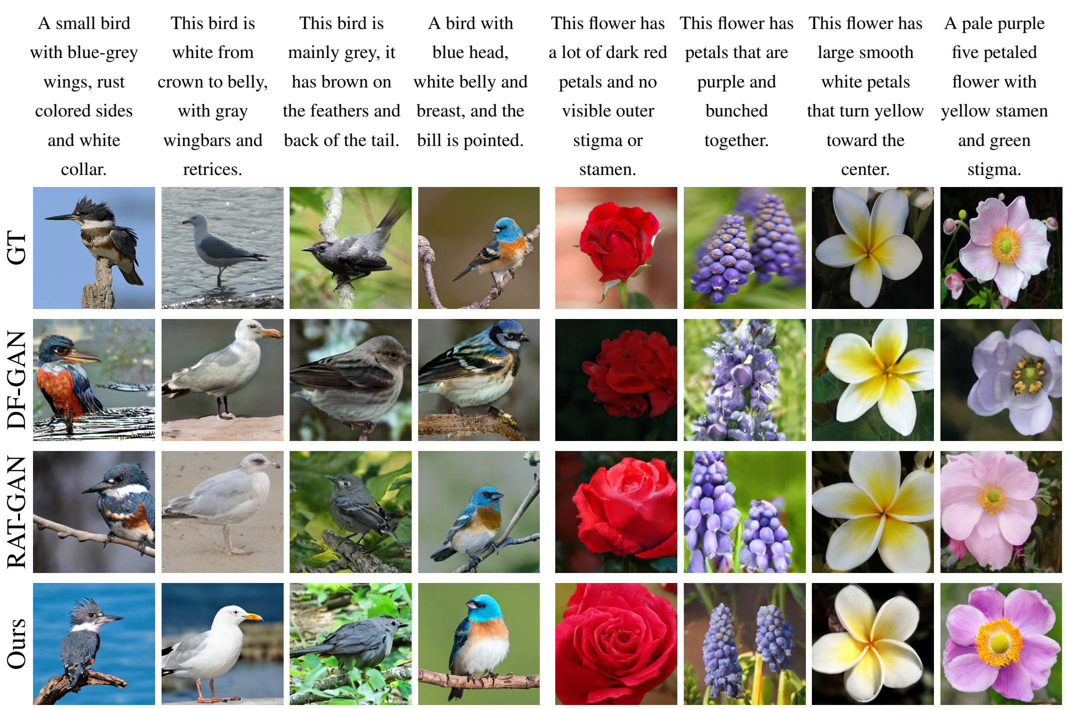

Figure 3: Qualitative comparison on the CUB and Oxford dataset. The input text descriptions are given in the first row and the corresponding generated images from different methods are shown in the same column. Best view in color and zoom in.

Datasets. We report results on the popular CUB, Oxford-102, and MS COCO datasets. The CUB dataset includes 200 categories with a total of 11,788 bird images, while the Oxford-102 dataset contains 102 categories with 8,189 flower images. Unlike the approaches taken in Reed et al. (2016a;b), we utilize the entire dataset for both training and testing. Each image is paired with 10 captions. To expand the original datasets, we collect 300,000 bird images and 130,000 flower images. The MS COCO dataset comprises 123,287 images, each with 5 sentence annotations. We use the official training split of COCO for training and the official validation split for testing. During mini-batch selection, a random image view (e.g., crop or flip) is chosen for one of the captions.

Web images. For the CUB and Oxford datasets, we collected 603,484 bird images and 331,602 flower images using search engines, utilizing fine-grained classification labels as search keywords. After removing detected outliers, we retained 399,246 bird images and 132,648 flower images. In the case of the COCO dataset, we gathered 770,059 daily images without applying any outlier detection, as the precise descriptions in COCO allow search engines to retrieve clean images effectively.

Table 1: Performance of IS and FID of StackGAN++, AttnGAN, SSGAN, DM-GAN, DTGAN, DF-GAN and our method on the CUB, Oxford and MS COCO datasets. The results are taken from the authors' own papers. The best results are in bold.

Methods	IS(Fine-tune) ↑		IS(ImageNet) ↑		FID(Fine-tune) ↓		FID(ImageNet) ↓		
	CUB	Oxford	CUB	Oxford	CUB	Oxford	CUB	Oxford	COCO
StackGAN++	4.04	3.26	4.04	3.26	23.96	48.68	15.30	32.33	81.59
AttnGAN	4.36	-	4.36	-	-	-	23.98	-	35.49
DAE-GAN	4.42	-	-	-	-	-	15.19	-	28.12
DM-GAN	4.75	-	-	-	-	-	16.09	-	32.64
DF-GAN	5.10	3.80	4.96	3.92	17.23	18.90	14.81	22.56	21.42
RAT-GAN	5.36	4.09	5.00	3.95	13.91	16.04	10.21	18.68	14.60
GALIP	-	-	-	-	-	-	10.05	-	5.85
VQ-Diffusion	-	-	-	-	-	-	10.32	14.10	13.86
U-ViT	-	-	-	-	-	-	-	-	5.45
Ours	**6.56**	**4.35**	**6.37**	**4.11**	**7.91**	**8.58**	**6.36**	**9.52**	**5.00**

Training details. The text encoder is a pre-trained CLIP text encoder with an output of size 512. The latent encoder and decoder is pre-trained by Stable Diffusion (Rombach et al., 2022). We have tried to pre-train new latent encoders on extrapolated data but the results are not satisfying. Adam optimizer is used to optimize the network with base learning rates of 0.0001 and weight decay of 0. The same as RAT-GAN, we used a mini-batch size of 24 to train the model. Most training and testing of our model are conducted on 2 RTX 3090 Ti and the detailed training consumption is listed in Table 3.

Evaluation metrics. We adopt the widely used Inception Score (IS) (Salimans et al., 2016) and Fréchet Inception Distance (FID) (Heusel et al., 2017) to quantify the performance. On the MS COCO dataset, an Inception-v3 network pre-trained on the ImageNet dataset is used to compute the KL-divergence between the conditional class distribution (generated images) and the marginal class distribution (real images). The presence of a large IS indicates that the generated images are of high quality. The FID computes the Fréchet Distance between the image feature distributions of the generated and real-world images. The image features are extracted by the same pre-trained Inception v3 network. A lower FID implies the generated images are closer to the real images. We only compare the FID on the COCO dataset. On the CUB and Oxford-102 dataset, pre-trained Inception models are fine-tuned on two fine-grained classification tasks (Zhang et al., 2019).

There are two conflicts in evaluation methods in previous works. First, some studies report Inception Score (IS) using the ImageNet Inception model, while others use a fine-tuned version. Second, some works evaluate using the entire training data, whereas others use only the test split. To address these inconsistencies, we report IS and FID using both Inception models and employ the same Inception model as DM-GAN for consistency. Additionally, to resolve conflicts related to data splits, we report the FID scores of our model and other re-implemented models using the entire dataset for both training and testing. According to results from RAT-GAN (Ye et al., 2023), training and testing on the full dataset typically yields the best FID scores. We will also release all evaluation codes on GitHub.

Compared models. We compare our model with recent state-of-the-art methods: Stack-GAN++ (Zhang et al., 2019), DM-GAN (Zhu et al., 2019), DF-GAN (Tao et al., 2022), DAE-GAN (Ruan et al., 2021), VQ-diffusion (Gu et al., 2022), AttnGAN (Xu et al., 2018), GALIP (Zhang & Schomaker, 2021),U-ViT (Bao et al., 2023), and RAT-GAN (Ye et al., 2023).

4.1 Comparisons with Others

Quantitative results. We present results for the CUB dataset of bird images, the Oxford-102 dataset of flower images, and the MS COCO dataset of common objects, as shown in Table 1. On the CUB dataset, our model achieve an IS score of 6.56 and an FID score of 6.36, outperforming all the previous models. For the Oxford dataset, we achieve an IS score of 4.35 and an FID score of 6.36, outperforming all the previous models. On the COCO dataset, our model achieves an

Table 2: Ablation studies on the CUB dataset. We utilize a NULL-guidance ratio of 1.5 during sampling. The FID score was employed to evaluate generation performance.

ID	Component				Extrapolation Quantity(k)					
	Cluster	Classification	RAT	NULL	0	50	100	200	300	400
0	-	-	-	-	16.74	-	-	-	-	30.78
1	✓	-	-	-	-	-	-	-	-	20.67
2	-	✓	-	-	-	-	-	-	-	12.45
3	✓	✓	-	-	-	-	-	-	-	9.87
4	✓	✓	✓	-	-	-	-	-	-	8.76
5	✓	✓	-	✓	-	-	-	-	-	7.65
6	✓	✓	✓	✓	-	9.56	7.34	6.87	6.54	**6.36**

FID score of 5.00 that is competitive with previous best result.Compared with VQ-Diffusion, our model uses less training data and achieve much better performance. This comparison reveals that pre-training on large datasets can be inefficient and lead to suboptimal results. Moreover, results in Table 1 reveal that Inception model pre-trained on ImageNet is less sensitive than fine-tuned on small datasets. Additionally, the Inception score on the Oxford dataset exceeds that of real images (4.10). Extensive results demonstrate the effectiveness and generalization ability of the proposed data extrapolation method.

Qualitative results. We present qualitative results for the CUB dataset of bird images and the Oxford-102 dataset of flower images. In Figure 3 , we compare the visualization results of DF-GAN, RAT-GAN, and our model. DF-GAN and RAT-GAN are previous state-of-the-art methods for text-to-image synthesis. On the CUB dataset, with more clear details such as feathers, eyes, and feet, our model clearly outperforms DF-GAN and RAT-GAN. Additionally, the background in our model's results is more coherent compared to RAT-GAN. On the Oxford dataset, our model exhibits better texture and more relevant colors than the others. With the proposed text extrapolation, RAT block, and null-guidance, our model demonstrates fewer distorted shapes and more relevant content compared to the other two models.

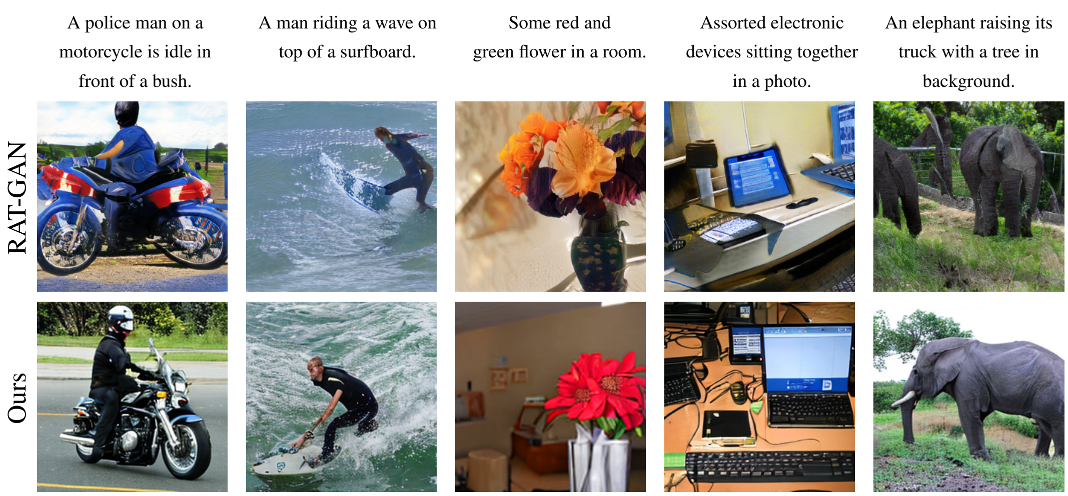

Figure 4: Qualitative comparison of our model with RAT-GAN on the COCO dataset.

The qualitative results for the COCO dataset are shown in Figure 4. The COCO dataset includes a wide variety of common objects, which makes it particularly susceptible to the long-tail problem (Chen et al., 2022). With additional training data obtained through extrapolation, our model generates more realistic objects compared to RAT-GAN. However, the collected 770,059 images are still insufficient to cover the entire distribution of images in COCO. As a result, the outputs from COCO are not as realistic as those from the CUB and Oxford datasets.

Table 3: Training consumption on the CUB, Oxford and COCO datasets. Fine-tuning is performed on the original dataset until the FID scores increase.

Dataset	Device	Original dataset	Extrapolated data	Fine-tuning
CUB	2 RTX 3090 Ti	5 days/1500 epochs	10 days/100 epochs	6 hours/50 epochs
Oxford	2 RTX 3090 Ti	5 days/1500 epochs	8 days/200 epochs	6 hours/50 epochs
COCO	2 RTX 4090	10 days/125 epochs	20 days/95 epochs	7 days/70 epochs

4.2 Ablation Studies

Analysis of outlier detectors. In Table 2, we present text-to-image results without cluster detector or classification detector. According to ID 0,1 and 2, the FID score without outlier detectors degrade severely because noisy images force the diffusion model to generate irrelevant objects. Although fine-tuning on small datasets could alleviate noise pollution but parameters also forget general knowledge at the same time. According to ID 2 and 3, classification detector performs better than cluster detector because it has utilized fine-grained classification labels.

Analysis of extrapolation quantity. More images generally lead to improved text-to-image results. however, this trend saturates around 100,000 images, after which the improvement in FID becomes less significant with more training samples. This phenomenon aligns with that diffusion models perform much better than GANs on the COCO dataset (84K images) but exhibit similar performance to GANs on the CUB and Oxford datasets(10K images). Furthermore, with transformers as core building blocks, GALIP performs similarly to previous models on the CUB dataset. This suggests that transformer architectures exacerbate the need for larger training datasets.

Analysis of NULL guidance. The performance of NULL guidance is influenced by both the NULL prompt and the guidance ratio. The results in Table 4 indicate that a NULL prompt reflecting the average meaning of the dataset achieves the best performance. Additionally, a suitable guidance ratio is crucial for optimal results, and we find that a ratio around 1.5 yields the best performance on the CUB and COCO datasets. However, on the Oxford dataset, NULL guidance improves the Inception Score from 4.10 to 4.35 but degrades the FID score from 9.52 to 11.07.

Table 4: The impact of various NULL prompts on FID scores in the CUB dataset.

NULL Prompts	Guidance Ratio 1.25	1.5	2.0
“Null”	7.23	7.16	7.68
“a picture”	6.89	6.54	7.14
“no description”	6.97	6.47	7.25
“a picture of bird”	6.46	6.36	6.86
“a picture of flower”	9.04	10.6	11.4
“we don’t know what it is”	8.98	9.35	9.94

Analysis of text injection. Text injection is crucial for text-to-image generation. As shown in ID 4 and 5 of Table 2, RAT significantly improves the FID score. Further experiments indicate that directly mixing text feature with time embedding results in an FID score of 25.41, which is much worse than 16.74 achieved by RAT. This suggests that time embedding provides information very different to text embedding. Additionally, incorporating a scaling operator into RAT can lead to model collapse, as information becomes highly compressed in latent space. Consequently, the mean value of the latent code becomes sensitive, and the scaling operation disrupts the information structure.

Table 5: Ablation studies on the MS COCO dataset. We adopt “A picture” as the NULL prompt.

Training data	FID score $\eta = 1.0$	$\eta = 1.5$	$\eta = 2.0$
COCO	11.89	7.99	8.43
Extrapolation	12.33	8.41	9.24
COCO-ft	8.45	5.00	5.56

Ablation studies on the MS COCO dataset. We conduct ablation studies on the MS COCO dataset, as presented in Table 5. The MS COCO dataset differs significantly from the CUB and Oxford datasets in terms of variety and image quantity. Experimental results demonstrate that linear

This flower is purple and white, and has petals that are bulb shaped and drooping downward.

Figure 5: Randomly generated images from the Oxford dataset. Best view in color and zoom in.

Model	FID	Type	Pre-training images	#Params
Parti (Yu et al., 2022)	**3.22**	Autoregressive	4.8B	20B
Make-A-Scene (Gafni et al., 2022)	7.55	Autoregressive	35M	4B
Re-Imagen (Chen et al., 2022)	5.25	Diffusion	50M	2.5B
VQ-Diffusion (Gu et al., 2022)	19.75	Diffusion	15M	370M
Ours	5.00	Diffusion	7M	464M

Table 6: Comparison of pre-training dataset and parameter quantity of different models on the MS COCO dataset. Parameters for text encoder, latent encoder and super resolution are not counted.

extrapolation and fine-tuning (5.00) outperform the original COCO dataset (7.99). However, unlike CUB and Oxford, fine-tuning on COCO requires much more time, as shown in Table 3. Additionally, we observe that early stopping is unnecessary for fine-tuning on the COCO dataset due to its larger image volume compared to CUB and Oxford.

In Table 6, we compare the pre-training dataset and model parameters with previous models on the MS COCO dataset. The compared models are all pre-trained on external datasets and fine-tuned on MS COCO dataset. Our result outperforms all previous models except for Parti but we use much less pre-training images and parameters than Parti. Moreover, our diffusion model is designed for small datasets and requires very few GPUs for training.

Diversity. To qualitatively evaluate the diversity of our proposed model, we generate random images conditioned on the same text description and different random noises. In Figure 5, we present 10 images generated from the same text. These images exhibit similar foreground elements while showcasing high diversity in spatial structure, demonstrating that our model effectively controls the image content.

5 Conclusion and Future work

In this paper, we propose a new data augmentation method for text-to-image generation using linear extrapolation. Specifically, we apply linear extrapolation only on text data, and new image data are retrieved from the internet by search engines. For the reliability of new text-image pairs, we design two outlier detectors to purify retrieved images. Based on extrapolation, we construct training samples dozens of times larger than the original dataset, resulting in a significant improvement in text-to-image performance. Moreover, we propose a NULL-condition guidance to refine the score estimation for text-to-image generation. This guidance is also applicable to existing text-to-image models without further training. In the future, linear extrapolation and NULL-condition guidance could be applied to tasks beyond text-to-image generation.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8861–8870, 2024.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

Fangxiang Feng, Tianrui Niu, Ruifan Li, and Xiaojie Wang. Modality disentangled discriminator for text-to-image synthesis. *IEEE Trans. Multim.*, 24:2112–2124, 2022.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pp. 89–106. Springer, 2022.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. Ieee, 2013.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 2020.

Xianxu Hou, Xiaokang Zhang, Yudong Li, and Linlin Shen. Textface: Text-to-style mapping based face generation and manipulation. *IEEE Transactions on Multimedia*, 2022.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Bowen Li, Philip HS Torr, and Thomas Lukasiewicz. Memory-driven text-to-image generation. *arXiv preprint arXiv:2208.07022*, 2022.

Vivian Liu, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 3dall-e: Integrating text-to-image ai in 3d design workflows. In *Proceedings of the 2023 ACM designing interactive systems conference*, pp. 1955–1977, 2023.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pp. 6448–6458. PMLR, 2020.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. URL `http://arxiv.org/abs/1411.1784`.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Jun Peng, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Knowledge-driven generative adversarial network for text-to-image synthesis. *IEEE Transactions on Multimedia*, 2021.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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.

Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *NIPS*, 2016a.

Scott E Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *international conference on machine learning*, pp. 1060–1069, 2016b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. DAE-GAN: dynamic aspect-aware GAN for text-to-image synthesis. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 13940–13949. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01370. URL `https://doi.org/10.1109/ICCV48922.2021.01370`.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pp. 2226–2234, 2016.

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pp. 30105–30118. PMLR, 2023.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 11895–11907, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR*. OpenReview.net, 2021.

Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. Cross-modal semantic matching generative adversarial networks for text-to-image synthesis. *IEEE Trans. Multim.*, 2022.

Ming Tao, Hao Tang, Fei Wu, Xiaoyuan Jing, Bing-Kun Bao, and Changsheng Xu. DF-GAN: A simple and effective baseline for text-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16494–16504. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01602. URL `https://doi.org/10.1109/CVPR52688.2022.01602`.

Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14214–14223, 2023.

Andrea Vedaldi and Andrew Zisserman. Vgg convolutional neural networks practical. *Department of Engineering Science, University of Oxford*, 66, 2016.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 2018.

Senmao Ye and Fei Liu. Score mismatching for generative modeling. *Neural Networks*, 175:106311, 2024.

Senmao Ye, Huan Wang, Mingkui Tan, and Fei Liu. Recurrent affine transformation for text-to-image synthesis. *IEEE Transactions on Multimedia*, 2023.

Ming-Yue Yin and Jian-Guang Li. A systematic review on digital human models in assembly process planning. *The International Journal of Advanced Manufacturing Technology*, 125(3):1037–1059, 2023.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1947–1962, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhenxing Zhang and Lambert Schomaker. DTGAN: dual attention generative adversarial networks for text-to-image generation. In *IJCNN*, pp. 1–8. IEEE, 2021.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, 2019.