RE-IMAGEN: RETRIEVAL-AUGMENTED TEXT-TO-IMAGE GENERATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Research on text-to-image generation has witnessed significant progress in generating diverse and photo-realistic images, driven by diffusion and auto-regressive models trained on large-scale image-text data. Though state-of-the-art models can generate high-quality images of common entities, they often have difficulty generating images of uncommon entities, such as 'Chortai (dog)' or 'Picarones (food)'. To tackle this issue, we present the Retrieval-Augmented Text-to-Image Generator (Re-Imagen), a generative model that uses retrieved information to produce high-fidelity and faithful images, even for rare or unseen entities. Given a text prompt, Re-Imagen accesses an external multi-modal knowledge base to retrieve relevant (image, text) pairs, and uses them as references to generate the image. With this retrieval step, Re-Imagen is augmented with the knowledge of high-level semantics and low-level visual details of the mentioned entities, and thus improves its accuracy in generating the entities' visual appearances. We train Re-Imagen on a constructed dataset containing (image, text, retrieval) triples to teach the model to ground on both text prompt and retrieval. Furthermore, we develop a new sampling strategy to interleave the classifier-free guidance for text and retrieval condition to balance the text and retrieval alignment. Re-Imagen achieves new SoTA FID results on two image generation benchmarks, such as COCO (*i.e.*, FID = 5.25) and WikiImage (*i.e.*, FID = 5.82) without fine-tuning. To further evaluate the capabilities of the model, we introduce EntityDrawBench, a new benchmark that evaluates image generation for diverse entities, from frequent to rare, across multiple visual domains. Human evaluation on EntityDrawBench shows that Re-Imagen performs on par with the best prior models in photo-realism, but with significantly better real-world faithfulness, especially on less frequent entities.

1 INTRODUCTION

Recent research efforts on conditional generative modeling, such as Imagen (Saharia et al., 2022), DALL·E 2 (Ramesh et al., 2022), and Parti (Yu et al., 2022), have advanced text-to-image generation to an unprecedented level, producing accurate, diverse, and even creative images from text prompts. These models leverage paired image-text data at Web scale (with hundreds of millions of training examples), and powerful backbone generative models, *i.e.*, autoregressive models (Van Den Oord et al., 2017; Ramesh et al., 2021; Yu et al., 2022), diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021), *etc.*, and generate highly realistic images. Studying these models' generation results, we discovered their outputs are surprisingly sensitive to the frequency of the entities (or objects) in the text prompts. In particular, when generating text prompts about frequent entities (see Appendix B), these models often generate realistic images, with faithful grounding to the entities' visual appearance. However, when generating from text prompts with less frequent entities, those models either hallucinate non-existent entities, or output related frequent entities (see Figure 1), failing to establish a connection between the generated image and the visual appearance of the mentioned entity. This key limitation can greatly harm the trustworthiness of text-to-image models in real-world applications and even raise ethical concerns.

In this paper, we propose a **Re**trieval-augmented Text-to-**Ima**ge **Gen**erator (Re-Imagen), which alleviates such limitations by searching for entity information in a multi-modal knowledge base, rather than attempting to memorize the appearance of rare entities. Specifically, we define our multi-modal knowledge base encodes the visual appearances and descriptions of entities with a collection of refer-

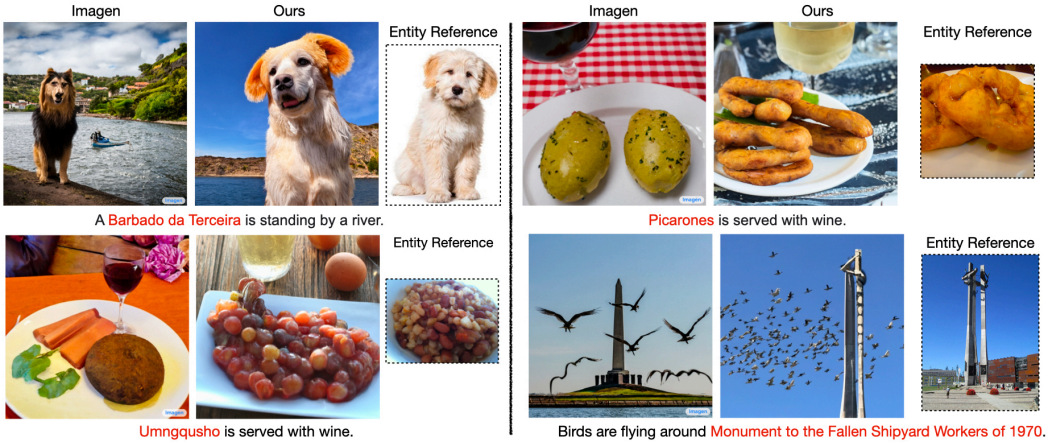

Figure 1: Comparison of images generated by Imagen and Re-Imagen on less frequent entities. We observe that Imagen hallucinates the entities while Re-Imagen maintains better faithfulness.

ence <image, text> pairs'. To use this resource, Re-Imagen first uses the input text prompt to retrieve the most relevant <image, text> pairs from the external multi-modal knowledge base, then uses the retrieved knowledge as model additional inputs to synthesize the target images. Consequently, the retrieved references provide knowledge regarding the semantic attributes and the concrete visual appearance of mentioned entities to guide Re-Imagen to paint the entities in the target images.

The backbone of Re-Imagen is a cascaded diffusion model (Ho et al., 2022), which contains three independent generation stages (implemented as U-Nets (Ronneberger et al., 2015)) to gradually produce high-resolution (*i.e.*, 1024×1024) images. In particular, we train Re-Imagen on a dataset constructed from the image-text dataset used by Imagen (Saharia et al., 2022), where each data instance is associated with the top-k nearest neighbors within the dataset, based on text-only BM25 score. The retrieved top-k <image, text> pairs will be used as a reference for the model attend to. During inference, we design an interleaved guidance schedule that switches between text guidance and retrieval guidance, which ensures both text alignment and entity alignment. We show some examples generated by Re-Imagen, and compare them against Imagen in Figure 1. We can qualitatively observe that the our images are more faithful to the appearance of the reference entity.

To further quantitatively evaluate Re-Imagen, we present *zero-shot text-to-image generation* results on two challenging datasets: COCO (Lin et al., 2014) and WikiImages (Chang et al., 2022)[1]. Re-Imagen uses an external non-overlapping image-text database as the knowledge base for retrieval and then grounds on the retrieval to synthesize the target image. We show that Re-Imagen achieves the state-of-the-art performance for text-to-image generation on COCO and WikiImages, measured in FID score Heusel et al. (2017), among non-fine-tuned models. For non-entity-centric dataset COCO, the perform gain is coming from biasing the model to generate images with similar styles as the retrieved in-domain images. For the entity-centric dataset WikiImages, the performance gain comes from grounding the generation on retrieved images containing similar entities. We further evaluate Re-Imagen on a more challenging benchmark — EntityDrawBench, to test the model's ability to generate a variety of infrequent entities (dogs, landmarks, foods) in different scenes. We compare Re-Imagen with Imagen (Saharia et al., 2022), DALL-E 2 (Ramesh et al., 2022) and StableDiffusion (Rombach et al., 2022) in terms of faithfulness and photorealism with human raters. We demonstrate that that Re-Imagen can reach around 71% on the faithfulness score, greatly improving from Imagine's score of 27% and DALL-E 2's score of 48% . Analysis shows that the improvement mostly comes from low-frequency entities.

To summarize, our key contributions are: (1) a novel retrieval-augmented text-to-image model Re-Imagen, which achieves SoTA FID scores on two dasets; (2) interleaved classifier-free guidance during sampling to ensure both text alignment and entity fidelity; and (3) We introduce EntityDrawBench and show that Re-Imagen can significantly improve faithfulness on less-frequent entities.

[1]The original WikiImages database contains (entity image, entity description) pairs. It was a crawled from Wikimedia Commons for visual question answering, and we repurpose it here for text-to-image generation.

2 RELATED WORK

Text-to-Image Diffusion Models There has been a wide-spread success (Ashual et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Nichol et al., 2021) in modeling text-to-image generation with diffusion models, which has outperformed GANs (Goodfellow et al., 2014) and auto-regressive Transformers (Ramesh et al., 2021) in photorealism and diversity (under similar model size), without training instability and mode collapsing issues. Among them, some recent large text-to-image models such as Imagen (Saharia et al., 2022), GLIDE (Nichol et al., 2021), and DALL-E2 (Ramesh et al., 2022) have demonstrated excellent generation from complex prompt inputs. These models achieve highly fine-grained control over the generated images with text inputs. However, they do not perform explicit grounding over external visual knowledge and are restricted to memorizing the visual appearance of every possible visual entity in their parameters. This makes it difficult for them to generalize to rare or even unseen entities. In contrast, Re-Imagen is designed to free the diffusion model from memorizing, as models are encouraged to retrieve semantic neighbors from the knowledge base and use retrievals as context to paint the image. Re-Imagen improves the grounding of the diffusion models to real-world knowledge and is therefore capable of faithful image synthesis.

Concurrent Work There are several concurrent works (Li et al., 2022; Blattmann et al., 2022; Ashual et al., 2022), that also leverage retrieval to improve diffusion models. RDM (Blattmann et al., 2022) is trained similarly to Re-Imagen, using examples and near neighbors, but the neighbors in RDM are selected using image features, and at inference time retrievals are replaced with user-chosen exemplars. RDM was shown to effectively transfer artistic style from exemplars to generated images. In contrast, our proposed Re-Imagen conditions on both text and multi-modal neighbors to generate the image, includes retrieval at inference time, and is demonstrated to improve performance on rare images (as well as more generally). KNN-Diffusion (Ashual et al., 2022) is more closely related work to us, as it also uses retrieval to the quality of generated images. However, KNN-Diffusion uses discrete image representations, while Re-Imagen uses the raw pixels, and Re-Imagen's retrieved neighbors can be <image, text> pairs, while KNN-Diffusion's are only images. Quantitatively, Re-Imagen outperforms KNN-Diffusion on the COCO dataset significantly.

Others Due to the space limit, we provide an additional literature review in the Appendix A.

3 MODEL

In this section, we discuss our proposed Re-Imagen in detail. We start with background knowledge, in the form of a brief overview of the cascaded diffusion models used by Imagen (Wang et al., 2022). Next, we describe the concrete technical details on how we incorporate multi-modal retrieval for Re-Imagen. Finally, we discuss training and the interleaved guidance sampling for Re-Imagen.

3.1 PRELIMINARIES

Diffusion Models Diffusion models (Sohl-Dickstein et al., 2015) are latent variable models, parameterized by θ, in the form of $p_\theta(\boldsymbol{x}_0) := \int p_\theta(\boldsymbol{x}_{0:T}) d\boldsymbol{x}_{1:T}$, where $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T$ are "noised" latent versions of the input image $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$. Note that the dimensionality of both latents and the image are the same throughout the entire process, with $\boldsymbol{x}_{0:T} \in \mathbb{R}^d$ and d equals the product of <height, width, # of channels>. The process that computes the posterior distribution $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$ is also called the forward (or diffusion) process, and is implemented as a predefined Markov chain that gradually adds Gaussian noise to the data according to a schedule β_t:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \qquad q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{1}$$

Diffusion models are trained to learn the image distribution by reversing the diffusion Markov chain. Theoretically, this reduces to learning to denoise $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ into $\boldsymbol{x}_0$, with a time re-weighted square error loss—see Ho et al. (2020) for the complete proof:

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}, t}[w_t \cdot ||\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \boldsymbol{x}_0||_2^2] \tag{2}$$

Here, the noised image is denoted as $\boldsymbol{x}_t := \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + (1-\bar{\alpha}_t)\boldsymbol{\epsilon}$, $\boldsymbol{x}_0$ is the ground-truth image, $\boldsymbol{c}$ is the condition, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the noise term, $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. To simplify

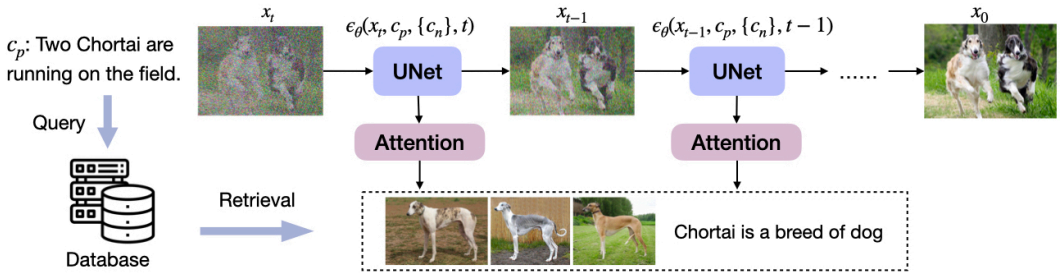

Figure 2: An illustration of the text-to-image generation pipeline in the 64× diffusion model. Specifically, Re-Imagen learns a UNet to iteratively predict $\boldsymbol{\epsilon}(\boldsymbol{x_t}, \boldsymbol{c}_n, \boldsymbol{c}_p, t)$ that denoises the image. ($\boldsymbol{c}_n$: a set of retrieved image-text pairs from the database; $\boldsymbol{c}_p$: input text prompt; t: current time-step)

notation, we will allow the condition $\boldsymbol{c}$ include multiple conditioning signals, such as text prompts $\boldsymbol{c}_p$, a low-resolution image input $\boldsymbol{c}_x$ (which is used in super-resolution), or retrieved neighboring images $\boldsymbol{c}_n$ (which are used in Re-Imagen). Imagen (Saharia et al., 2022) uses a U-Net (Ronneberger et al., 2015) to implement $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t)$. The U-Net represents the reversed noise generator as follows:

$$\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) := (\boldsymbol{x}_t - (1 - \bar{\alpha}_t)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t))/\sqrt{\bar{\alpha}_t} \tag{3}$$

During the training, we randomly sample $t \sim \mathcal{U}([0, 1])$ and image $\boldsymbol{x}_0$ from the dataset $\mathcal{D}$, and minimize the difference between $\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c})$ and $\boldsymbol{x}_0$ according to Equation 2. At the inference time, the diffusion model uses DDPM (Ho et al., 2020) to sample recursively as follows:

$$\boldsymbol{x}_{t-1} = \frac{\sqrt{\alpha_t}\beta_t}{1 - \bar{\alpha}_t}\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \sqrt{\alpha_t}}\boldsymbol{x}_t + \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t\boldsymbol{\epsilon} \tag{4}$$

The model sets $\boldsymbol{x}_T$ as a Gaussian noise with T denoting the total number of diffusion steps, and then keep sampling in reverse until step $T = 0$, i.e. $\boldsymbol{x}_T \to \boldsymbol{x}_{T-1} \to \cdots$, to reach the final image $\hat{\boldsymbol{x}}_0$.

For better generation efficiency, cascaded diffusion models (Ho et al., 2022; Ramesh et al., 2022; Saharia et al., 2022) use three separate diffusion models to generate high-resolution images gradually, going from low resolution to high resolution. The three models 64× model, 256× super-resolution model and 1024× super-resolution model gradually increase the model resolution to 1024×1024.

Classifier-free Guidance Ho & Salimans (2021) first proposed classifier-free guidance to trade off diversity and sample quality. This sampling strategy has been widely used due to its simplicity. In particular, Imagen (Saharia et al., 2022) adopts an adjusted ϵ-prediction as follows:

$$\hat{\boldsymbol{\epsilon}} = w \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t) - (w - 1) \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{5}$$

where w is the guidance weight. The unconditional ϵ-prediction $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is calculated by dropping the condition, i.e. the text prompt.

3.2 Generating Image with Multi-Modal Knowledge

Similar to Imagen (Saharia et al., 2022), Re-Imagen is a cascaded diffusion model, consisting of 64×, 256×, and 1024× diffusion models. However, Re-Imagen augments the diffusion model with the new capability of leveraging multimodal 'knowledge' from the external database, thus freeing the model from memorizing the appearance of rare entities. For brevity (and concreteness) we present below a high-level overview of the 64× model: the others are similar.

Main Idea As shown in Figure 2, during the denoising process, Re-Imagen conditions its generation result not only on the text prompt $\boldsymbol{c}_p$ (and also with $\boldsymbol{c}_x$ for super-resolution), but on the neighbors $\boldsymbol{c}_n$ that were retrieved from the external knowledge base. Here, the text prompt $\boldsymbol{c}_p \in \mathbb{R}^{n \times d}$ is represented using a T5 embedding (Raffel et al., 2020), with n being the text length and d being the embedding dimension. Meanwhile, the top-k neighbors $\boldsymbol{c}_n := [\text{<image, text>}_1, \cdots, \text{<image, text>}_k]$ are retrieved from external knowledge base $\mathcal{B}$, using the input prompt p as the query and a retrieval similarity function $\gamma(p, \mathcal{B})$. We experimented with two different choices for the similarity function: maximum inner product scores for BM25 (Robertson et al., 2009) and CLIP (Radford et al., 2021).

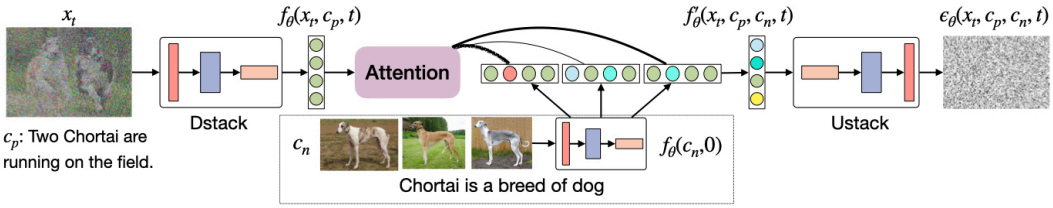

Figure 3: The detailed architecture of our model. The retrieved neighbors are first encoded using the DStack encoder and then used to augment the intermediate representation of the denoising image (via cross-attention). The augmented representation is fed to the UStack to predict the noise.

Model Architecture We show the architecture of our model in Figure 3, where we decompose the UNet into the downsampling encoder (DStack) and the upsampling decoder (UStack). Specifically, the DStack takes an image, a text, and a time step as the input, and generates a feature map, which is denoted as $f_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, t) \in \mathbb{R}^{F \times F \times d}$, with F denoting the feature map width and d denoting the hidden dimension. We share the same DStack encoder when we encode the retrieved <image, text> pairs (with t set to zero) which produces a set of feature maps $f_\theta(\boldsymbol{c}_n, 0) \in \mathbb{R}^{K \times F \times F \times d}$. We then use a multi-head attention module (Vaswani et al., 2017) to extract the most relevant information to produce a new feature map $f'_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, \boldsymbol{c}_n, t) = Attn(f_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, t), f_\theta(\boldsymbol{c}_n, 0))$. The upsampling stack decoder then predicts the noise term $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, \boldsymbol{c}_n, t)$ and uses it to compute $\hat{\boldsymbol{x}}_\theta$ with Equation 3, which is either used for regression during training or DDPM sampling.

Model Training In order to train Re-Imagen, we construct a new dataset KNN-ImageText based on the 50M ImageText-dataset used in Imagen. There are two motivations for selecting this dataset. (1) the dataset contains many similar photos regarding specific entities, which is extremely helpful for obtaining similar neighbors, and (2) the dataset is highly sanitized with fewer unethical or harmful images. For each instance in the 50M ImageText-dataset, we search over the same dataset with text-to-text BM25 similarity to find the top-2 neighbors as $\boldsymbol{c}_n$ (excluding the query instance). We experimented with both CLIP and BM25 similarity scores, and retrieval was implemented with ScaNN (Guo et al., 2020). We train Re-Imagen on the KNN-ImageText by minimizing the loss function of Equation 2. During training, we also randomly drop the text and neighbor conditions independently with 10% chance. Such random dropping will help the model learn the marginalized noise term $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, t)$ and $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_n, t)$, which will be used for the classifier-free guidance.

Interleaved Classifier-free Guidance Different from existing diffusion models, our model needs to deal with more than one condition, *i.e.*, text prompts $\boldsymbol{c}_t$ and retrieved neighbors $\boldsymbol{c}_n$, which allows new options for incorporating guidance. In particular, Re-Imagen could use classifier-free guidance by subtracting the unconditioned ϵ-predictions, or either of the two partially conditioned ϵ-predictions. Empirically, we observed that subtracting unconditioned ϵ-predictions (the standard classifier-free guidance of Figure 3.1) often leads to an undesired imbalance, where the outputs are either dominated by the text condition or the neighbor condition. Hence, we designed an interleaved guidance schedule that balances the two conditions. Formally, we define the two adjusted ϵ-predictions as:

$$\begin{aligned} \hat{\boldsymbol{\epsilon}}_p &= w_p \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, \boldsymbol{c}_n, t) - (w_p - 1) \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_n, t) \\ \hat{\boldsymbol{\epsilon}}_n &= w_n \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, \boldsymbol{c}_n, t) - (w_n - 1) \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, t) \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{\epsilon}}_p$ and $\hat{\boldsymbol{\epsilon}}_n$ are the text-enhanced and neighbor-enhanced ϵ-predictions, respectfully. Here, w_p is the text guidance weight and w_n is the neighbor guidance weight. We then interleave the two guidance predictions by a certain predefined ratio η. Specifically, at each guidance step, we sample a [0, 1]-uniform random number R, and $R < \eta$, we use $\hat{\boldsymbol{\epsilon}}_p$, and otherwise $\hat{\boldsymbol{\epsilon}}_n$. We can adjust η to balance the faithfulness w.r.t text description or the retrieved image-text pairs. In EntityDrawBench experiment, we found that $\eta = 0.55$ can lead to better quality.

4 Experiments

Re-Imagen consists of three submodels: a 2.5B 64×64 text-to-image model, a 750M 256×256 super-resolution model and a 400M 1024×1024 super-resolution model. We finetune these models on the constructed KNN-ImageText dataset. We evaluate the model under two settings: (1) automatic evaluation on COCO and WikiImages dataset, to measure the model's general performance

Model	# of Params	FID-30K	Zero-shot FID-30K
GLIDE (Nichol et al., 2021)	5B	-	12.24
DALL-E 2 (Ramesh et al., 2022)	∼5B	-	10.39
VQ-Diffusion (Gu et al., 2022)	0.4B	-	19.75
KNN-Diffusion (Ashual et al., 2022)	0.8B	-	16.66
Stable-Diffusion (Rombach et al., 2022)	1B	-	12.63
Imagen (Saharia et al., 2022)	3B	-	7.27
Make-A-Scene (Gafni et al., 2022)	4B	7.55	11.84
Parti (Yu et al., 2022)	20B	**3.22**	7.23
Re-Imagen (γ=BM25; $\mathcal{B}$=COCO; k=2)	3.6B	**5.25**†	-
Re-Imagen (γ=CLIP; $\mathcal{B}$=COCO; k=2)	3.6B	5.29†	-
Re-Imagen (γ=BM25; $\mathcal{B}$=ImageText; k=2)	3.6B	-	7.02
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=2)	3.6B	-	6.88

Table 1: MS-COCO results for zero-shot text-to-image generation. We use a guidance weight of 1.25 for the 64× diffusion model and 5 for our 256× super-resolution model. (†: Re-Imagen *is not fine-tuned* on the COCO data—it only uses it as the knowledge base for retrieval.)

to generate photorealistic images, and (2) human evaluation on the newly introduced EntityDraw-Bench, to measure the model's capability to generate long-tail entities.

Training and Evaluation details The fine-tuning was run for 200K steps on 64 TPU-v4 chips and completed within two days. We use Adafactor for the 64× model and Adam for the 256× super-resolution model with a learning rate of 1e-4. We set the number of neighbors k=2 and set γ=BM25 during training. For the image-text database $\mathcal{B}$, we consider three different variants: (1) the COCO/WikiImages training set, which contains non-overlapping small-scale in-domain image-text pairs, (2) the ImageText dataset containing 50M <image, caption> pairs, and (3) the LAION dataset (Schuhmann et al., 2021) containing 400M <image, text> crawled pairs. Since indexing ImageText and LAION with CLIP encodings is expensive, we only considered the BM25 retriever for these databases. For the COCO/WikiImages training set, we used both BM25 and CLIP.

4.1 Evaluation on COCO and WikiImages

In these two experiments, we used the standard non-interleaved classifier-free guidance (Figure 3.1) with T=1000 steps for both the 64× diffusion model and 256× super-resolution model. The guidance weight w for the 64× model is swept over [1.0, 1.25, 1.5, 1.75, 2.0], while the 256×256 super-resolution models' guidance weight w is swept over [1.0, 5.0, 8.0, 10.0]. We select the guidance w with the best FID score, which is reported in Table 1. We also demonstrate examples in Figure 4.

COCO Results COCO is the most widely-used benchmark for text-to-image generation models. Although COCO does not contain many rare entities, it does contain unusual combinations of common entities, so it is plausible that retrieval augmentation could also help for some challenging text prompts. We adopt FID (Heusel et al., 2017) score to measure image quality. Following the previous literature, we randomly sample 30K prompts from the validation set as input to the model. The generated images are compared with the reference images from the full validation set (42K). We list the results in two columns: FID-30K denotes the model with access to the COCO train set (either to fine-tune or retrieve from), while Zero-shot FID-30K does not have access to any COCO data.

Re-Imagen (with the COCO database) can achieve a significant gain on FID-30K without fine-tuning: roughly a 2.0 absolute FID improvement over Imagen. The performance is even better than fine-tuned Make-A-Scene (Gafni et al., 2022), but slightly worse than fine-tuned 20B Parti. In contrast, Re-Imagen retrieving from out-of-domain databases (LAION) achieves less gain, but still obtains a 0.4 FID improvement over Imagen. Re-Imagen outperforms KNN-Diffusion, another retrieval-augmented diffusion model, by a large margin.

Since COCO does not contain infrequent entities, 'entity knowledge' is not important. In contrast, retrieving from the training set can provide useful 'style knowledge' for the model to ground on. Re-Imagen is able to adapt the generated images to the same style of the COCO distribution, it can achieve a much better FID score. As can be seen in the upper part of from Figure 4, Re-Imagen with

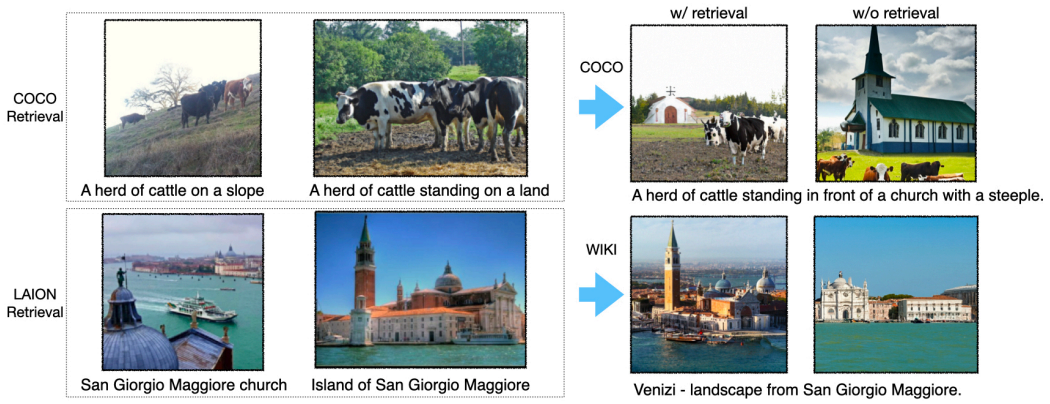

Figure 4: The retrieved top-2 neighbors of COCO and WikiImages and model generation.

Model	# of Params	FID-30K	Zero-shot FID-20K
Stable-Diffusion (Rombach et al., 2022)	1B	-	7.50
Imagen (Saharia et al., 2022)	3B	-	6.44
Re-Imagen (γ=BM25; $\mathcal{B}$=WikiImages; k=2)	3.6B	5.88	-
Re-Imagen (γ=CLIP; $\mathcal{B}$=WikiImages; k=2)	3.6B	5.85	-
Re-Imagen (γ=BM25; $\mathcal{B}$=ImageText; k=2)	3.6B	-	6.04
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=1)	3.6B	-	5.94
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=2)	3.6B	-	5.82
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=3)	3.6B	-	**5.80**

Table 2: WikiImages results for zero-shot text-to-image generation. We use a guidance weight of 1.5 for the 64$\times$ diffusion model and 5 for our 256$\times$ super-resolution model.

retrieval generates images of the same style as COCO, while without retrieval, the output is still high quality, but the style is less similar to COCO.

WikiImages Results WikiImages is constructed based on the multimodal corpus provided in WebQA (Chang et al., 2022), which consists of <image, text> pairs crawled from Wikimedia Commons[2]. We filtered the original corpus to remove noisy data (see the AppendixC), which leads to a total of 320K examples. We randomly sample 22K as our validation set to perform zero-shot evaluation, we further sample 20K prompts from the dataset as the input. Similar to the previous experiment, we also adopt the guidance weight schedule as before and evaluate 256$\times$256 images. We report our experimental results in Table 2 and mainly compare with Imagen and Stable-Diffusion.

From Table 2, we found that using out-of-domain LAION-400M as the database actually achieves better performance than using in-domain WikiImages as the database. Unlike COCO, WikiImages contains mostly entity-focused images, thus the importance of finding relevant entities is the database is more important than distilling the styles from the training set—and since the scale of LAION-400M is 100x larger than WikiImages-300K, the chance of retrieving related entities is much higher, which leads to better performance. One example is depicted in the lower part of Figure 4, where the LAION retrieval finds 'Island of San Giorgio Maggiore', which helps the model generate the classical Renaissance-style church. When generating without retrieval, the model is not able to generate the specific church. This indicates the importance of having relevant entities in the retrievals for WikiImages dataset and also explains why LAION database achieves the best results. We also present more examples from WikiImages in the Appendix C.

4.2 Entity Focused Evaluation on EntityDrawBench

Dataset Construction We introduce EntityDrawBench to evaluate the model's capability to generate diverse sets of entities in different visual scenes. Specifically, we pick three types of visual entities (dog breeds, landmarks, and foods) from Wikipedia Commons and Google Landmarks to

[2] https://commons.wikimedia.org/wiki/Main_Page

Model	Faithfulness				Photorealism
	Dogs	Foods	Landmarks	All	All
Imagen	0.28 ± 0.02	0.26 ± 0.02	0.27 ± 0.02	0.27	**0.98**
DALL-E 2	0.60 ± 0.02	0.47 ± 0.02	0.36 ± 0.04	0.48	**0.98**
Stable-Diffusion	0.16 ± 0.02	0.24 ± 0.04	0.12 ± 0.06	0.17	0.92
Re-Imagen	**0.68** ± 0.04	**0.70** ± 0.02	**0.74** ± 0.04	**0.71**	0.97

Table 3: Human evaluation results for different models on different types of entities.

construct our prompts. In total, we collect 150 entity-centric prompts for evaluation. These prompts are mostly unique and we cannot find corresponding images with Google Image Search. More construction details are in Appendix D.

We use the prompt as the input and its corresponding image-text pairs as the 'retrieval' for Re-Imagen, to generate four 1024×1024 images. For the other models, we feed the prompts directly also to generate four images. We will pick the best image of these four samples to rate its Photorealism and Faithfulness. For photorealism, we assign 1 if the image is moderately realistic without noticeable artifacts, otherwise, we assign a score of 0. For the faithfulness measure, we assign 1 if the image is faithful to both the entity source and the text description, otherwise, we assign 0.

Experimental Results We use the proposed interleaved classifier-free guidance (subsection 3.2) for the 64× diffusion model, which runs for 256 diffusion steps under a strong guidance weight of w=30 for both text and neighbor conditions. For the 256× and 1024× resolution models, we use a constant guidance weight of 5.0 and 3.0, respectively, with 128 and 32 diffusion steps. The inference speed is 30-40 secs for 4 images on 4 TPU-v4 chips. We demonstrate our human evaluation results for faithfulness and photorealism in Table 3.

We can observe that Re-Imagen can in general achieve much higher faithfulness than the existing models while maintaining similar photorealism scores. When comparing with our backbone Imagen, we see the faithfulness score jumps from 27% to 71%, which indicates that our model is paying attention to the retrieved knowledge and assimilating it into the generation process.

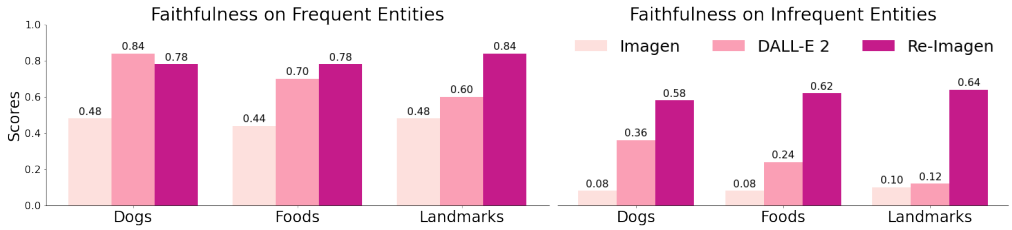

Figure 5: The human evaluation scores for both frequent and infrequent entities.

We further partition the entities into 'frequent' and 'infrequent' categories based on their frequency (top 50% as 'frequent') in Imagen's training corpus. We plot faithfulness score for 'frequent' and 'infrequent' separately in Figure 5 . We can see that our model is less sensitive to the frequency of the input entities than the other models, with only a 10-20% drop on infrequent entities. In contrast, both Imagen and DALL-E 2 drop by 40%-50% on infrequent entities. This study reflects the effectiveness of text-to-image generation models on long-tail entities.

Comparison to Other Models We demonstrate some examples from different models in Figure 6. As can be seen, the images generated from Re-Imagen strike a good balance between text alignment and entity fidelity. Unlike image editing to perform in-place modification, Re-Imagen can transform the neighbor entities both geometrically and semantically according to the text guidance. As a concrete example, Re-Imagen generates the *Braque Saint-Germain* (2nd row in Figure 6) on the grass, in a different viewpoint from to the reference image.

Text and Entity Faithfulness Trade-offs In our experiments, we found that there is a trade-off between faithefulness to the text prompt and faithfulness to the retrieved entity images. Based on Equation 6, by adjusting η, *i.e.*the proportion of $\hat{\epsilon}_p$ and $\hat{\epsilon}_n$ in the sampling schedule, we can control Re-Imagen so as to generate images that explore this tradeoff: decreasing η will increase

Figure 6: None-cherry picked examples from EntityDrawBench for different models.

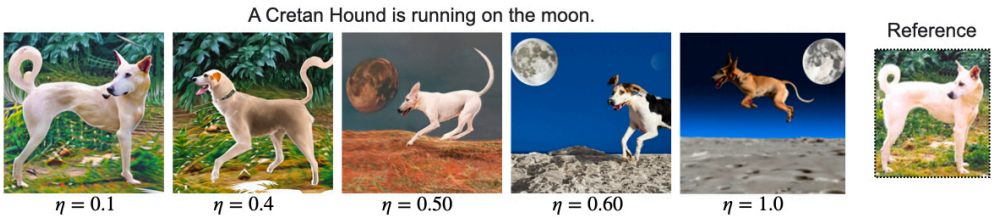

Figure 7: Ablation study of interleaved guidance ratio η to show the trade-off.

the entity's image faithfulness but decrease the text alignment. In contrast, increasing the value η will increase the text alignment and decrease the similarity to retrieved image. We demonstrate this in Figure 7. With small η, the model ignores the text description and simply copies the retrieved image, while with large η, the model reverts to standard Imagen, without using to the input image much. We found that having η around 0.5 is usually a 'sweet spot' that balances both conditions.

5 Conclusions and Limitations

We present Re-Imagen, a retrieval-augmented diffusion model, and demonstrate its effectiveness in generating realistic and faithful images. We exhibit such advantages not only through automatic FID measures on standard benchmarks (*i.e.*, COCO and WikiImage) but also through human evaluation on the newly introduced EntityDrawBench. We further demonstrate that our model is particularly effective in generating an image from text that mentions rare entities.

Re-Imagen still suffers from well-known issues in text-to-image generation, which we review below in section 5. In addition, Re-Imagen also has some unique limitations due to the retrieval-augmented modeling. First, because Re-Imagen is sensitive the to retrieved image-text pairs it is conditioned on, when the retrieved image is of low-quality, there will be a negative influence on the generated image. Second, Re-Imagen sometimes still fail to ground on the retrieved entities when the entity's visual appearance is out of the generation space. Third, we noticed that the super-resolution model is less effective, and frequently misses low-level texture details of the visual entities. In future work, we plan to further investigate the above limitations and address them.

ETHICS STATEMENT

Strong text-to-image generation models, *i.e.*, Imagen (Saharia et al., 2022) and Parti (Yu et al., 2022), raise ethical challenges along dimensions such as the *social bias*. Re-Imagen is exposed to the same challenges, as we employed Web-scale datasets that are similar to the prior models.

The retrieval-augmented modeling techniques of Re-Imagen has substantially improved the controllability and attribution of the generated image. Like many basic research topics, this additional control could be used for beneficial or harmful purposes. One obvious danger is that Re-Imagen (or similar models) could be used for malicious purposes like spreading misinformation, *e.g.,* by producing realistic images of specific people in misleading visual contexts. On the other side, additional control has many potential benefits. One general benefit is that Re-Imagen can reduce hallucination and increase the faithfulness of the generated image to the users' intent. Another benefit is that the ability to work with tail entities makes the model more useful for minorities and other users in smaller communities: for example, Re-Imagen is more effective at generating images of landmarks famous in smaller communities or cultures, and generating images of indigenous foods and cultural artifacts. We argue that this model can help decrease the frequency-caused bias in current neural network based AI systems.

Considering such potential threats to the public, we have currently decide not to release the code or a public demo. In future work, we will explore a framework for responsible use that balances the value of external auditing of research with the risks of unrestricted open access, allowing this work to be used in a safe and beneficial way.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18511–18521, 2022.

Oron Ashual, Shelly Sheynin, 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.

Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2021.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16495–16504, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

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. *arXiv preprint arXiv:2203.13131*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

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.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pp. 3887–3896. PMLR, 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 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. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23: 47–1, 2022.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

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

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6959–6969, 2022.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. 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*, 2022.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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.

Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, number FZJ-2022-00923. Jülich Supercomputing Center, 2021.

Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Retrievalfuse: Neural 3d scene reconstruction with a database. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12568–12577, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11379–11388, 2022.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2575–2584, 2020.

Rui Xu, Minghao Guo, Jiaqi Wang, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Texture memory-augmented deep patch-based image inpainting. *IEEE Transactions on Image Processing*, 30:9112–9124, 2021.

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*, 2022.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pp. 592–608. Springer, 2020.

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pp. 597–613. Springer, 2016.