GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

Alex Nichol [* 1] **Prafulla Dhariwal** [* 1] **Aditya Ramesh** [* 1] **Pranav Shyam** [1] **Pamela Mishkin** [1] **Bob McGrew** [1]
Ilya Sutskever [1] **Mark Chen** [1]

Abstract

Diffusion models have recently been shown to generate high-quality synthetic images, especially when paired with a guidance technique to trade off diversity for fidelity. We explore diffusion models for the problem of text-conditional image synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance. We find that the latter is preferred by human evaluators for both photorealism and caption similarity, and often produces photorealistic samples. Samples from a 3.5 billion parameter text-conditional diffusion model using classifier-free guidance are favored by human evaluators to those from DALL-E, even when the latter uses expensive CLIP reranking. Additionally, we find that our models can be fine-tuned to perform image inpainting, enabling powerful text-driven image editing. We train a smaller model on a filtered dataset and release the code and weights at `https://github.com/openai/glide-text2im`.

1. Introduction

Images, such as illustrations, paintings, and photographs, can often be easily described using text, but can require specialized skills and hours of labor to create. Therefore, a tool capable of generating realistic images from natural language can empower humans to create rich and diverse visual content with unprecedented ease. The ability to edit images using natural language further allows for iterative refinement and fine-grained control, both of which are critical for real world applications.

Recent text-conditional image models are capable of synthesizing images from free-form text prompts, and can compose unrelated objects in semantically plausible ways (Xu et al., 2017; Zhu et al., 2019; Tao et al., 2020; Ramesh et al., 2021; Zhang et al., 2021). However, they are not yet able to generate photorealistic images that capture all aspects of their corresponding text prompts.

On the other hand, unconditional image models can synthesize photorealistic images (Brock et al., 2018; Karras et al., 2019a;b; Razavi et al., 2019), sometimes with enough fidelity that humans can't distinguish them from real images (Zhou et al., 2019). Within this line of research, diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2020b) have emerged as a promising family of generative models, achieving state-of-the-art sample quality on a number of image generation benchmarks (Ho et al., 2020; Dhariwal & Nichol, 2021; Ho et al., 2021).

To achieve photorealism in the class-conditional setting, Dhariwal & Nichol (2021) augmented diffusion models with *classifier guidance*, a technique which allows diffusion models to condition on a classifier's labels. The classifier is first trained on noised images, and during the diffusion sampling process, gradients from the classifier are used to guide the sample towards the label. Ho & Salimans (2021) achieved similar results without a separately trained classifier through the use of *classifier-free guidance*, a form of guidance that interpolates between predictions from a diffusion model with and without labels.

Motivated by the ability of guided diffusion models to generate photorealistic samples and the ability of text-to-image models to handle free-form prompts, we apply guided diffusion to the problem of text-conditional image synthesis. First, we train a 3.5 billion parameter diffusion model that uses a text encoder to condition on natural language descriptions. Next, we compare two techniques for guiding diffusion models towards text prompts: CLIP guidance and classifier-free guidance. Using human and automated evaluations, we find that classifier-free guidance yields higher-quality images. We find that samples from our model generated with classifier-free guidance are both photorealistic and reflect a wide breadth of world knowledge. When evaluated by human judges, our samples are preferred to those from
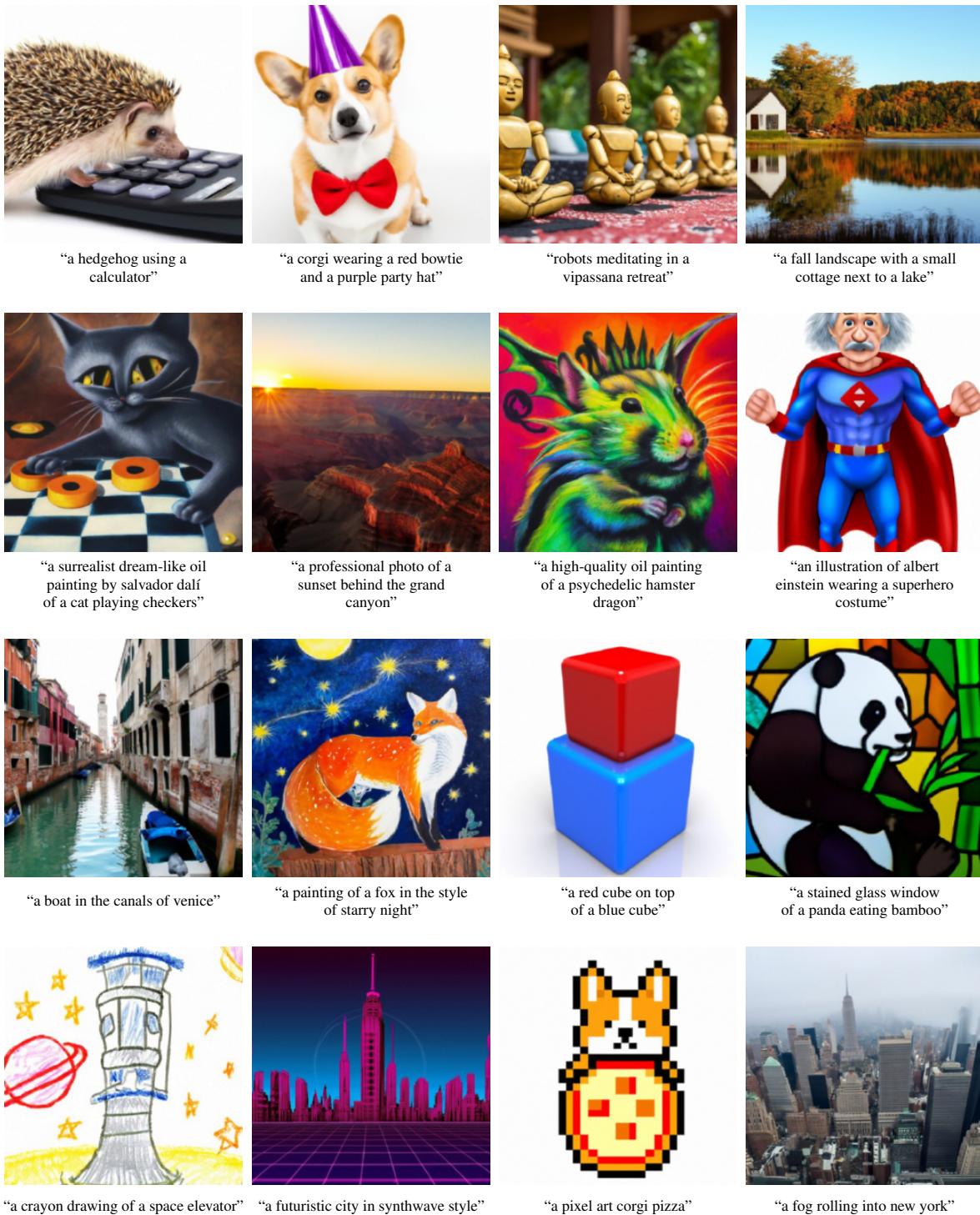
"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

"robots meditating in a vipassana retreat"

"a fall landscape with a small cottage next to a lake"

"a surrealist dream-like oil painting by salvador dalí of a cat playing checkers"

"a professional photo of a sunset behind the grand canyon"

"a high-quality oil painting of a psychedelic hamster dragon"

"an illustration of albert einstein wearing a superhero costume"

"a boat in the canals of venice"

"a painting of a fox in the style of starry night"

"a red cube on top of a blue cube"

"a stained glass window of a panda eating bamboo"

"a crayon drawing of a space elevator"

"a futuristic city in synthwave style"

"a pixel art corgi pizza"

"a fog rolling into new york"

Figure 1. Selected samples from GLIDE using classifier-free guidance. We observe that our model can produce photorealistic images with shadows and reflections, can compose multiple concepts in the correct way, and can produce artistic renderings of novel concepts. For random sample grids, see Figure 17 and 18.

DALL-E (Ramesh et al., 2021) 87% of the time when evaluated for photorealism, and 69% of the time when evaluated for caption similarity.

While our model can render a wide variety of text prompts zero-shot, it can have difficulty producing realistic images for complex prompts. Therefore, we provide our model with editing capabilities in addition to zero-shot generation,
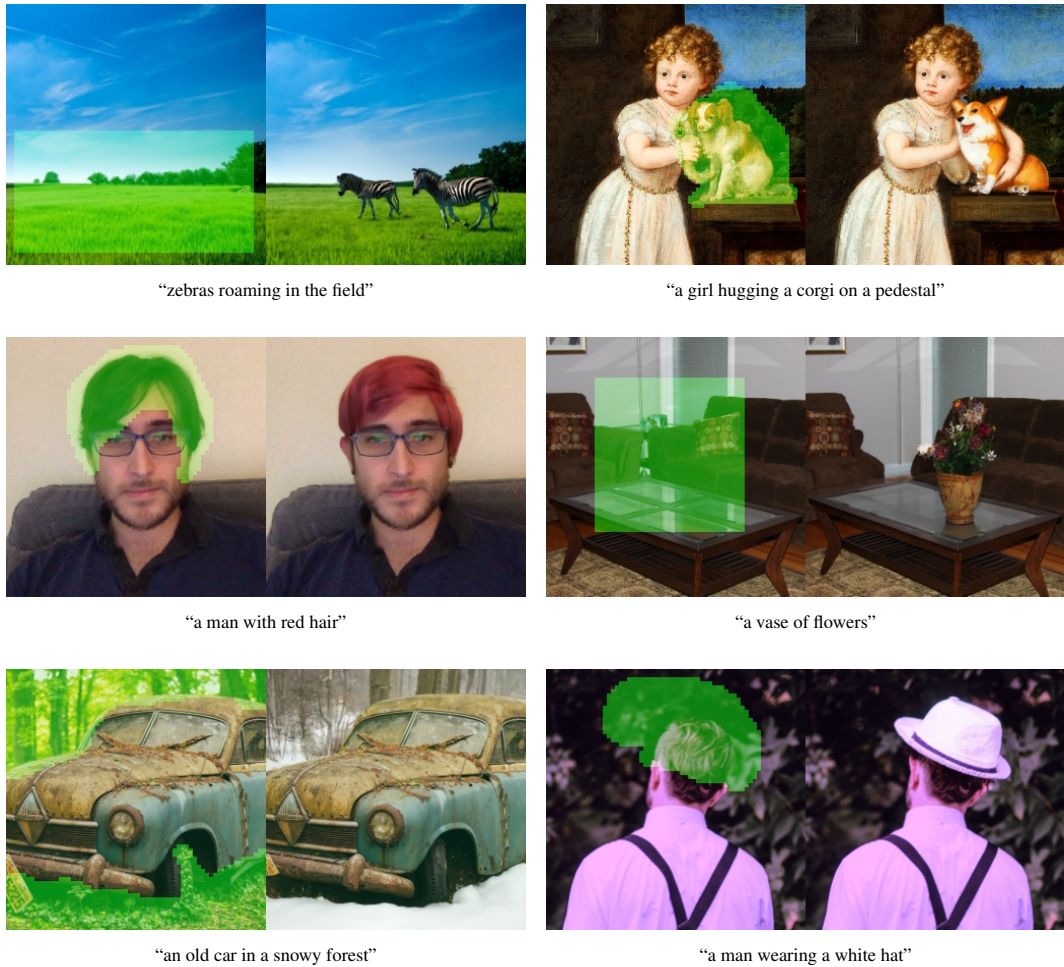
"zebras roaming in the field"

"a girl hugging a corgi on a pedestal"

"a man with red hair"

"a vase of flowers"

"an old car in a snowy forest"

"a man wearing a white hat"

Figure 2. Text-conditional image inpainting examples from GLIDE. The green region is erased, and the model fills it in conditioned on the given prompt. Our model is able to match the style and lighting of the surrounding context to produce a realistic completion.

which allows humans to iteratively improve model samples until they match more complex prompts. Specifically, we fine-tune our model to perform image inpainting, finding that it is capable of making realistic edits to existing images using natural language prompts. Edits produced by the model match the style and lighting of the surrounding context, including convincing shadows and reflections. Future applications of these models could potentially aid humans in creating compelling custom images with unprecedented speed and ease.

We observe that our resulting model can significantly reduce the effort required to produce convincing disinformation or Deepfakes. To safeguard against these use cases while aiding future research, we release a smaller set of models trained on filtered datasets (Appendix E).

We refer to our system as GLIDE, which stands for **G**uided **L**anguage to **I**mage **D**iffusion for Generation and **E**diting. We refer to our small filtered model as GLIDE (filtered).

2. Background

In the following sections, we outline the components of the final models we will evaluate: diffusion, classifier-free guidance, and CLIP guidance.

2.1. Diffusion Models

We consider the Gaussian diffusion models introduced by Sohl-Dickstein et al. (2015) and improved by Song & Ermon (2020b); Ho et al. (2020). Given a sample from the data distribution $x_0 \sim q(x_0)$, we produce a Markov chain of latent variables $x_1, ..., x_T$ by progressively adding Gaussian noise to the sample:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathcal{I})$$

If the magnitude $1 - \alpha_t$ of the noise added at each step is small enough, the posterior $q(x_{t-1}|x_t)$ is well-approximated by a diagonal Gaussian. Furthermore, if the magnitude $1 - \alpha_1...\alpha_T$ of the total noise added through-

"a cozy living room" | "a painting of a corgi on the wall above a couch" | "a round coffee table in front of a couch" | "a vase of flowers on a coffee table" | "a couch in the corner of a room"

Figure 3. Iteratively creating a complex scene using GLIDE. First, we generate an image for the prompt "a cozy living room", then use the shown inpainting masks and follow-up text prompts to add a painting to the wall, a coffee table, and a vase of flowers on the coffee table, and finally to move the wall up to the couch.

"a corgi wearing a bow tie and a birthday hat"

"a fire in the background"

Figure 4. Examples of text-conditional SDEdit (Meng et al., 2021) with GLIDE, where the user combines a sketch with a text caption to do more controlled modifications to an image.

out the chain is large enough, x_T is well approximated by $\mathcal{N}(0, \mathcal{I})$. These properties suggest learning a model $p_\theta(x_{t-1}|x_t)$ to approximate the true posterior:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(\mu_\theta(x_t), \Sigma_\theta(x_t))$$

which can be used to produce samples $x_0 \sim p_\theta(x_0)$ by starting with Gaussian noise $x_T \sim \mathcal{N}(0, \mathcal{I})$ and gradually reducing the noise in a sequence of steps $x_{T-1}, x_{T-2}, ..., x_0$.

While there exists a tractable variational lower-bound (VLB) on $\log p_\theta(x_0)$, better results arise from optimizing a surrogate objective which re-weighs the terms of the VLB. To compute this surrogate objective, we generate samples $x_t \sim q(x_t|x_0)$ by applying Gaussian noise ϵ to to x_0, then train a model ϵ_θ to predict the added noise using a standard mean-squared error loss:

$$L_{\text{simple}} := E_{t \sim [1,T], x_0 \sim q(x_0), \epsilon \sim \mathcal{N}(0, \mathbf{I})}[||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

Ho et al. (2020) show how to derive $\mu_\theta(x_t)$ from $\epsilon_\theta(x_t, t)$,

and fix Σ_θ to a constant. They also show the equivalence to previous denoising score-matching based models (Song & Ermon, 2020b;a), with the score function $\nabla_{x_t} \log p(x_t) \propto \epsilon_\theta(x_t, t)$. In a follow-up work, Nichol & Dhariwal (2021) present a strategy for learning Σ_θ, which enables the model to produce high quality samples with fewer diffusion steps. We adopt this technique in training the models in this paper.

Diffusion models have also been successfully applied to image super-resolution (Nichol & Dhariwal, 2021; Saharia et al., 2021b). Following the standard formulation of diffusion, high-resolution images y_0 are progressively noised in a sequence of steps. However, $p_\theta(y_{t-1}|y_t, x)$ additionally conditions on the downsampled input x, which is provided to the model by concatenating x (after bicubic upsampling) along the channel dimension. Results from these models outperform prior methods on FID, IS, and in human comparison scores.

2.2. Guided Diffusion

Dhariwal & Nichol (2021) find that samples from class-conditional diffusion models can often be improved with classifier guidance, where a class-conditional diffusion model with mean $\mu_\theta(x_t|y)$ and variance $\Sigma_\theta(x_t|y)$ is additively perturbed by the gradient of the log-probability $\log p_\phi(y|x_t)$ of a target class y predicted by a classifier. The resulting new perturbed mean $\hat{\mu}_\theta(x_t|y)$ is given by

$$\hat{\mu}_\theta(x_t|y) = \mu_\theta(x_t|y) + s \cdot \Sigma_\theta(x_t|y) \nabla_{x_t} \log p_\phi(y|x_t)$$

The coefficient s is called the guidance scale, and Dhariwal & Nichol (2021) find that increasing s improves sample quality at the cost of diversity.

2.3. Classifier-free guidance

Ho & Salimans (2021) recently proposed classifier-free guidance, a technique for guiding diffusion models that does not require a separate classifier model to be trained. For classifier-free guidance, the label y in a class-conditional

diffusion model $\epsilon_\theta(x_t|y)$ is replaced with a null label $\emptyset$ with a fixed probability during training. During sampling, the output of the model is extrapolated further in the direction of $\epsilon_\theta(x_t|y)$ and away from $\epsilon_\theta(x_t|\emptyset)$ as follows:

$$\hat{\epsilon}_\theta(x_t|y) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset))$$

Here $s \geq 1$ is the guidance scale. This functional form is inspired by the implicit classifier

$$p^i(y|x_t) \propto \frac{p(x_t|y)}{p(x_t)}$$

whose gradient can be written in terms of the true scores ϵ^*

$$\nabla_{x_t} \log p^i(y|x_t) \propto \nabla_{x_t} \log p(x_t|y) - \nabla_{x_t} \log p(x_t)$$
$$\propto \epsilon^*(x_t|y) - \epsilon^*(x_t)$$

To implement classifier-free guidance with generic text prompts, we sometimes replace text captions with an empty sequence (which we also refer to as $\emptyset$) during training. We then guide towards the caption c using the modified prediction $\hat{\epsilon}$:

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\emptyset))$$

Classifier-free guidance has two appealing properties. First, it allows a single model to leverage its own knowledge during guidance, rather than relying on the knowledge of a separate (and sometimes smaller) classification model. Second, it simplifies guidance when conditioning on information that is difficult to predict with a classifier (such as text).

2.4. CLIP Guidance

Radford et al. (2021) introduced CLIP as a scalable approach for learning joint representations between text and images. CLIP model consists of two models: an image encoder $f(x)$ and a caption encoder $g(c)$. During training, batches of (x, c) pairs are sampled from a large dataset, and the model optimizes a contrastive cross-entropy loss that encourages a high dot-product $f(x) \cdot g(c)$ if the image x is paired with the given caption c, or a low dot-product if the image and caption correspond to different pairs in the training data.

Since CLIP provides a score of how close an image is to a caption, several works have used it to steer generative models like GANs towards a user-defined text caption (Galatolo et al., 2021; Patashnik et al., 2021; Murdock, 2021; Gal et al., 2021). To apply the same idea to diffusion models, we can replace the classifier with a CLIP model in classifier guidance. In particular, we perturb the reverse-process mean with the gradient of the dot product of the image and caption encodings with respect to the image:

$$\hat{\mu}_\theta(x_t|c) = \mu_\theta(x_t|c) + s \cdot \Sigma_\theta(x_t|c) \nabla_{x_t} (f(x_t) \cdot g(c))$$

Similar to classifier guidance, we must train CLIP on noised images x_t to obtain the correct gradient in the reverse process. Throughout our experiments, we use CLIP models that were explicitly trained to be noise-aware, which we refer to as noised CLIP models.

Prior work Crowson (2021a;b) has shown that the public CLIP models, which have not been trained on noised images, can still be used to guide diffusion models. In Appendix C, we show that our noised CLIP guidance performs favorably to this approach without requiring additional tricks like data augmentation or perceptual losses. We hypothesize that guiding using the public CLIP model adversely impacts sample quality because the noised intermediate images encountered during sampling are out-of-distribution for the model.

3. Related Work

Many works have approached the problem of text-conditional image generation. Xu et al. (2017); Zhu et al. (2019); Tao et al. (2020); Zhang et al. (2021); Ye et al. (2021) train GANs with text-conditioning using publicly available image captioning datasets. Ramesh et al. (2021) synthesize images conditioned on text by building on the approach of van den Oord et al. (2017), wherein an autoregressive generative model is trained on top of discrete latent codes. Concurrently with our work, Gu et al. (2021) train text-conditional discrete diffusion models on top of discrete latent codes, finding that the resulting system can produce competitive image samples.

Several works have explored image inpainting with diffusion models. Meng et al. (2021) finds that diffusion models can not only inpaint regions of an image, but can do so conditioned on a rough sketch (or set of colors) for the image. Saharia et al. (2021a) finds that, when trained directly on the inpainting task, diffusion models can smoothly inpaint regions of an image without edge artifacts.

CLIP has previously been used to guide image generation. Galatolo et al. (2021); Patashnik et al. (2021); Murdock (2021); Gal et al. (2021) use CLIP to guide GAN generation towards text prompts. The online AI-generated art community has produced promising early results using unnoised CLIP-guided diffusion (Crowson, 2021a;b). Kim & Ye (2021) edits images using text prompts by fine-tuning a diffusion model to target a CLIP loss while reconstructing the original image's DDIM (Song et al., 2020a) latent. Zhou et al. (2021) trains GAN models conditioned on perturbed CLIP image embeddings, resulting in a model which can condition images on CLIP text embeddings. None of these works explore noised CLIP models, and often rely on data augmentations and perceptual losses as a result.

Several works have explored text-based image editing.

Zhang et al. (2020) propose a dual attention mechanism for using text embeddings to inpaint missing regions of an image. Stap et al. (2020) propose a method for editing images of faces using feature vectors grounded in text. Bau et al. (2021) pair CLIP with state-of-the-art GAN models to inpaint images using text targets. Concurrently with our work, Avrahami et al. (2021) use CLIP-guided diffusion to inpaint regions of images conditioned on text.

4. Training

For our main experiments, we train a 3.5 billion parameter text-conditional diffusion model at 64×64 resolution, and another 1.5 billion parameter text-conditional upsampling diffusion model to increase the resolution to 256×256. For CLIP guidance, we also train a noised 64×64 ViT-L CLIP model (Dosovitskiy et al., 2020).

4.1. Text-Conditional Diffusion Models

We adopt the ADM model architecture proposed by Dhariwal & Nichol (2021), but augment it with text conditioning information. For each noised image x_t and corresponding text caption c, our model predicts $p(x_{t-1}|x_t, c)$. To condition on the text, we first encode it into a sequence of K tokens, and feed these tokens into a Transformer model (Vaswani et al., 2017). The output of this transformer is used in two ways: first, the final token embedding is used in place of a class embedding in the ADM model; second, the last layer of token embeddings (a sequence of K feature vectors) is separately projected to the dimensionality of each attention layer throughout the ADM model, and then concatenated to the attention context at each layer. We train our model on the same dataset as DALL-E (Ramesh et al., 2021) and with roughly the same training compute as DALL-E. For training and hyperparameter details, see Appendix B.1.

4.2. Fine-tuning for classifier-free guidance

After the initial training run, we fine-tuned our base model to support unconditional image generation. This training procedure is exactly like pre-training, except 20% of text token sequences are replaced with the empty sequence. This way, the model retains its ability to generate text-conditional outputs, but can also generate images unconditionally.

4.3. Image Inpainting

Most previous work that uses diffusion models for inpainting has not trained diffusion models explicitly for this task (Sohl-Dickstein et al., 2015; Song et al., 2020b; Meng et al., 2021). In particular, diffusion model inpainting can be performed by sampling from the diffusion model as usual, but replacing the known region of the image with a sample

from $q(x_t|x_0)$ after each sampling step. This has the disadvantage that the model cannot see the entire context during the sampling process (only a noised version of it), occasionally resulting in undesired edge artifacts in our early experiments.

To achieve better results, we explicitly fine-tune our model to perform inpainting, similar to Saharia et al. (2021a). During fine-tuning, we randomly mask regions of the input images, and train the model on the same objective while conditioning it on the masked versions of the clean ground-truth images and the mask. To do this, we modify the model architecture by adding four new input channels: a second set of RGB channels to provide the masked ground-truth image, and a mask channel to provide the inpainting mask. We initialize the corresponding input weights for these new channels to zero before fine-tuning. For the upsampling model, we always provide the full low-resolution image, but only provide the unmasked region of the high-resolution image.

4.4. Noised CLIP models

To better match the classifier guidance technique from Dhariwal & Nichol (2021), we train noised CLIP models with an image encoder $f(x_t, t)$ that receives noised images x_t and is otherwise trained with the same objective as the original CLIP model. We train these models at 64×64 resolution with the same noise schedule as our base model.

5. Results

5.1. Qualitative Results

In Figure 1, we observe that GLIDE with classifier-free guidance is capable of generalizing to a wide variety of prompts. The model often generates realistic shadows and reflections, as well as high-quality textures. It is also capable of producing illustrations in various styles, such as the style of a particular artist or painting, or in general styles like pixel art. Finally, the model is able to compose several concepts (e.g. a corgi, bowtie, and birthday hat), all while binding attributes (e.g. colors) to these objects.

In Figure 5, we compare our model to previous state-of-the-art text-conditional image generation models on MS-COCO captions, finding that our model produces more realistic images without CLIP reranking or cherry-picking. We also see that classifier-free guidance produces more realistic images than CLIP guidance. We further show our model compares favorably against unnoised CLIP guidance in Appendix C, and evaluate the impact of model size in Appendix G.

On the inpainting task, we find that GLIDE can realistically modify existing images using text prompts, inserting new objects, shadows and reflections when necessary (Figure 2).
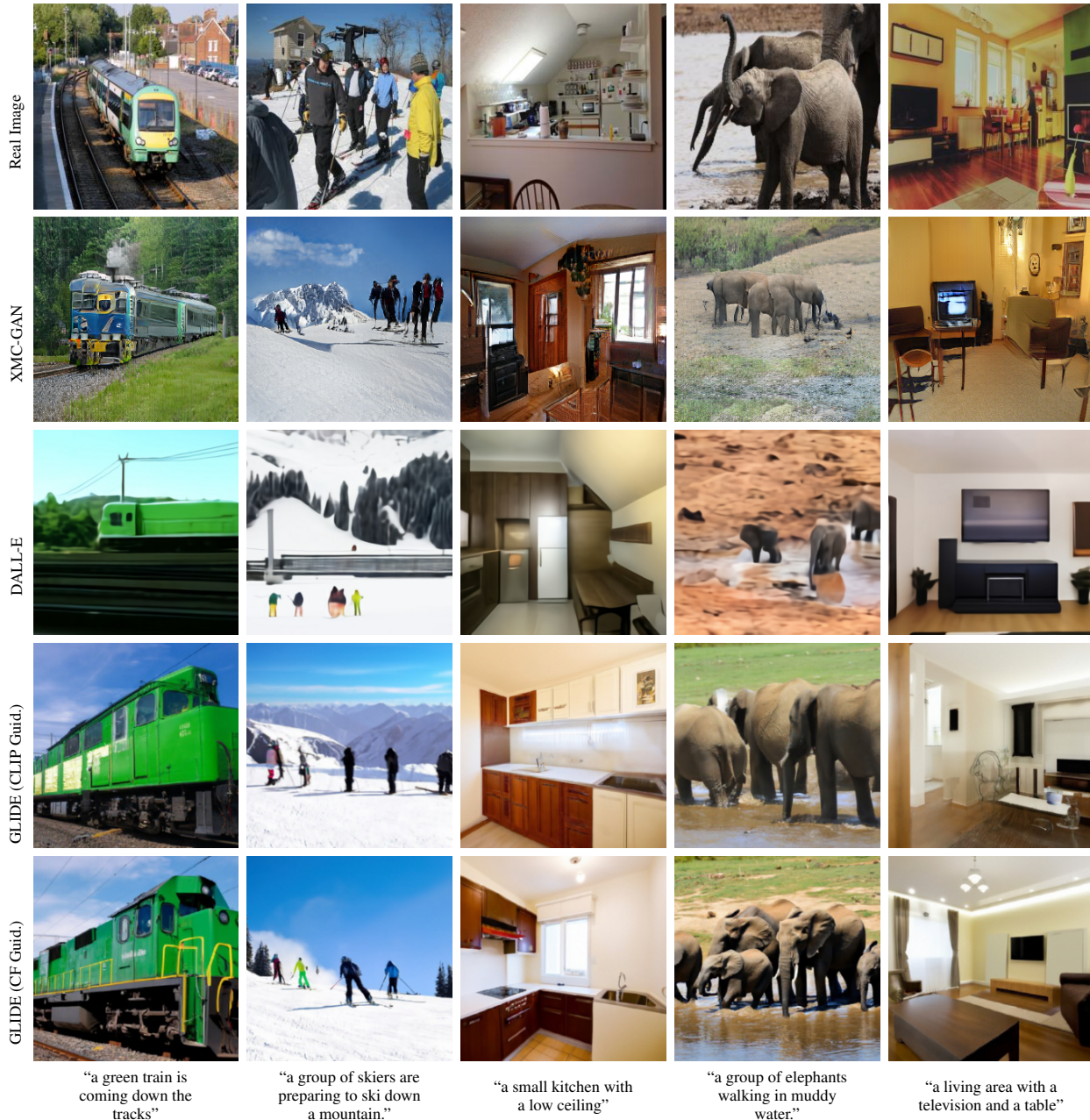
Figure 5. Random image samples on MS-COCO prompts. For XMC-GAN, we take samples from Zhang et al. (2021). For DALL-E, we generate samples at temperature 0.85 and select the best of 256 using CLIP reranking. For GLIDE, we use CLIP guidance with scale 2.0 and classifier-free guidance with scale 3.0. We do not perform any CLIP reranking or cherry-picking for GLIDE.

The model can even match styles when editing objects into paintings. In Figure 3 we show how we can use GLIDE iteratively to produce a complex scene using a zero-shot generation followed by a series of inpainting edits. We also experiment with SDEdit (Meng et al., 2021) in Figure 4, finding that our model is capable of turning sketches into realistic image edits. In Appendix D, we compare our model to previous inpainting models, showing our model produces more realistic completions.

5.2. Quantitative Results

We first evaluate the difference between classifier-free guidance and CLIP guidance by looking at the Pareto frontier of the quality-fidelity trade-off. In Figure 6 we evaluate both approaches on zero-shot MS-COCO generation at 64×64 resolution. We look at Precision/Recall (Kynkäänniemi et al., 2019), FID (Heusel et al., 2017), Inception Score (Salimans et al., 2016), and CLIP score (Radford et al., 2021). We define the CLIP score as
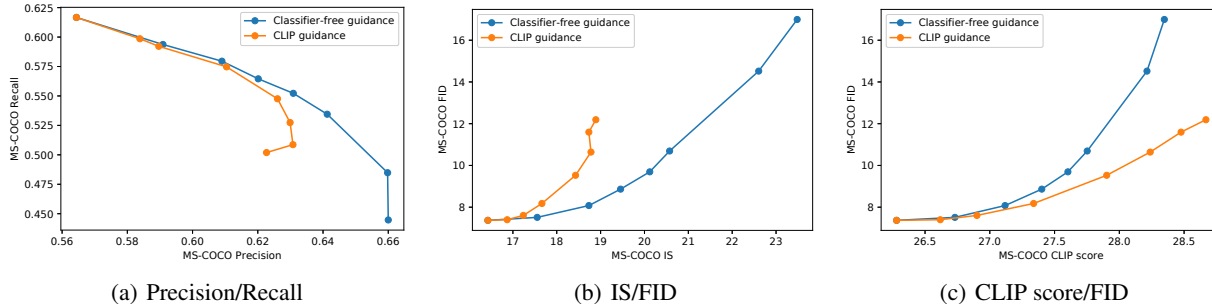
(a) Precision/Recall   (b) IS/FID   (c) CLIP score/FID

Figure 6. Comparing the diversity-fidelity trade-off of classifier-free guidance and CLIP guidance on MS-COCO 64×64.
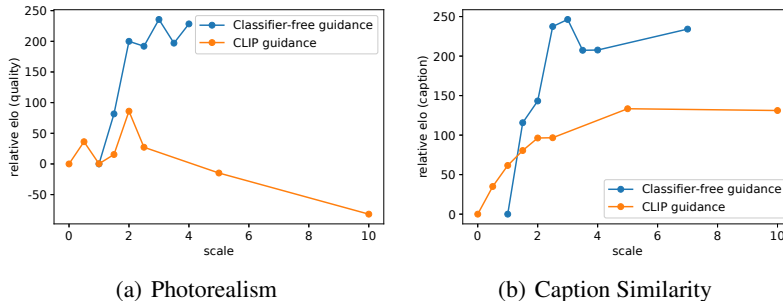

(a) Photorealism   (b) Caption Similarity

Figure 7. Elo scores from human evaluations for finding the optimal guidance scales for classifier-free guidance and CLIP guidance. The classifier-free guidance and CLIP guidance comparisons were performed separately, but can be super-imposed onto the same graph my normalizing for the Elo score of unguided sampling.

$\mathbb{E}[s(f(\text{image}) \cdot g(\text{caption}))]$ where the expectation is taken over the batch of samples and s is the CLIP logit scale. As we increase both guidance scales, we observe a clean trade-off in FID vs. IS, Precision vs. Recall, and CLIP score vs. FID. In the former two curves, we find that classifier-free guidance is (nearly) Pareto optimal. We see the exact opposite trend when plotting CLIP score against FID; in particular, CLIP guidance seems to be able to boost CLIP score much more than classifier-free guidance.

We hypothesize that CLIP guidance is finding adversarial examples for the evaluation CLIP model, rather than actually outperforming classifier-free guidance when it comes to matching the prompt. To verify this hypothesis, we employed human evaluators to judge the sample quality of generated images. In this setup, human evaluators are presented with two 256×256 images and must choose which sample either 1) better matches a given caption, or 2) looks more photorealistic. The human evaluator may also indicate that neither image is significantly better than the other, in which case half of a win is assigned to both models. Using our human evaluation protocol, we first sweep over guidance scales for both approaches separately (Figure 7), then compare the two methods with the best scales from the previous stage (Table 1). We find that humans disagree with CLIP score, finding classifier-free guidance to yield higher-quality samples that agree more with the prompt.

Table 1. Elo scores resulting from a human evaluation of unguided diffusion sampling, classifier-free guidance, and CLIP guidance on MS-COCO validation prompts at 256×256 resolution. For classifier-free guidance, we use scale 3.0, and for CLIP guidance scale 2.0. See Appendix A.1 for details on Elo score computation.

Guidance	Photorealism	Caption
Unguided	-88.6	-106.2
CLIP guidance	-73.2	29.3
Classifier-free guidance	**82.7**	**110.9**

We also compare GLIDE with other text-conditional generative image models. We find in Table 2 that our model obtains competitive FID on MS-COCO without ever explicitly training on this dataset. We also compute FID against a subset of the MS-COCO validation set that has been purged of all images similar to images in our training set, as done by Ramesh et al. (2021). This reduces the validation batch by 21%. We find that our FID increases slightly from 12.24 to 12.89 in this case, which could largely be explained by the change in FID bias when using a smaller reference batch.

Finally, we compare GLIDE against DALL-E using our human evaluation protocol (Table 3). Note that GLIDE was trained with roughly the same training compute as DALL-E but with a much smaller model (3.5 billion vs. 12 billion parameters). It also requires less sampling latency and no

Table 2. Comparison of FID on MS-COCO 256×256. Like previous work, we sample 30k captions for our models, and compare against the entire validation set. For our model, we report numbers for classifier-free guidance with scale 1.5, as it yields the best FID.

Model	FID	Zero-shot FID
AttnGAN (Xu et al., 2017)	35.49	
DM-GAN (Zhu et al., 2019)	32.64	
DF-GAN (Tao et al., 2020)	21.42	
DM-GAN + CL (Ye et al., 2021)	20.79	
XMC-GAN (Zhang et al., 2021)	9.33	
LAFITE (Zhou et al., 2021)	**8.12**	
DALL-E (Ramesh et al., 2021)		~ 28
LAFITE (Zhou et al., 2021)		26.94
GLIDE		**12.24**
GLIDE (Validation filtered)		**12.89**

Table 3. Human evaluation results comparing GLIDE to DALL-E. We report win probabilities of our model for both photorealism and caption similarity. In the final row, we apply the dVAE used by DALL-E to the outputs of GLIDE.

	DALL-E Temp.	Photo-realism	Caption Similarity
No reranking	1.0	91%	83%
	0.85	84%	80%
DALL-E reranked	1.0	89%	71%
	0.85	87%	69%
DALL-E reranked + GLIDE blurred	1.0	72%	63%
	0.85	66%	61%

CLIP reranking. We perform three sets of comparisons between DALL-E and GLIDE: first, we compare both models without CLIP reranking; second, we use CLIP reranking only for DALL-E; finally, we use CLIP reranking for DALL-E and run GLIDE samples through DALL-E's discrete VAE. We do all evals using two temperatures for the DALL-E model. Our model is preferred by the human evaluators in all settings, even in the configurations that heavily favor DALL-E by allowing it to use a much larger amount of test-time compute (through CLIP reranking) while reducing GLIDE sample quality (through VAE blurring). For sample grids from DALL-E with CLIP reranking and GLIDE with various guidance strategies, see Appendix H.

6. Conclusion and Future work

We have shown that text-conditional diffusion models are capable of producing realistic synthetic images, especially when paired with classifier-free guidance. While our model can often compose disparate concepts in complex ways, it sometimes fails to capture certain prompts which describe highly unusual objects or scenarios. In Figure 8, we provide some examples of these failure cases. Also, our unoptimized

"an illustration of a cat that has eight legs"

"a bicycle that has continuous tracks instead of wheels"

"a mouse hunting a lion"

"a car with triangular wheels"

Figure 8. Failure cases of GLIDE when prompted for certain unusual objects or scenarios.

model takes 15 seconds to sample one image on a single A100 GPU. This is much slower than sampling for related GAN methods, which produce images in a single forward pass and are thus more favorable for use in real-time applications. Capturing out of distribution prompts and sampling fast remain important directions for future work.

7. Acknowledgements

References

Avrahami, O., Lischinski, D., and Fried, O. Blended diffusion for text-driven editing of natural images. *arXiv:2111.14818*, 2021.

Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., and Torralba, A. Paint by word. *arXiv:2103.10951*, 2021.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv:1809.11096*, 2018.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional

accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Crowson, K. Clip guided diffusion hq 256x256. https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj, 2021a.

Crowson, K. Clip guided diffusion 512x512, secondary model method. https://twitter.com/RiversHaveWings/status/1462859669454536711, 2021b.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *arXiv:2105.05233*, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

Gal, R., Patashnik, O., Maron, H., Chechik, G., and Cohen-Or, D. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv:2108.00946*, 2021.

Galatolo, F. A., Cimino, M. G. C. A., and Vaglini, G. Generating images from caption and vice versa via clip-guided generative latent space search. *arXiv:2102.01645*, 2021.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis. *arXiv:2111.14822*, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv:2006.11239*, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *arXiv:2106.15282*, 2021.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *arXiv:arXiv:1812.04948*, 2019a.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. *arXiv:1912.04958*, 2019b.

Kim, G. and Ye, J. C. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv:2110.02711*, 2021.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. *arXiv:1904.06991*, 2019.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv:2108.01073*, 2021.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. Mixed precision training. *arXiv:1710.03740*, 2017.

Murdock, R. The big sleep. https://twitter.com/advadnoun/status/1351038053033406468, 2021.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models. *arXiv:2102.09672*, 2021.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. *arXiv:2103.17249*, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *arXiv:2103.00020*, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. *arXiv:2102.12092*, 2021.

Razavi, A., van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with VQ-VAE-2. *arXiv:1906.00446*, 2019.

Saharia, C., Chan, W., Chang, H., Lee, C. A., Ho, J., Salimans, T., Fleet, D. J., and Norouzi, M. Palette: Image-to-image diffusion models. *arXiv:2111.05826*, 2021a.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *arXiv:arXiv:2104.07636*, 2021b.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *arXiv:1606.03498*, 2016.

Santurkar, S., Tsipras, D., Tran, B., Ilyas, A., Engstrom, L., and Madry, A. Image synthesis with a single (robust) classifier. *arXiv:1906.09453*, 2019.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv:1503.03585*, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020a.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *arXiv:2006.09011*, 2020a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *arXiv:arXiv:1907.05600*, 2020b.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, 2020b.

Stap, D., Bleeker, M., Ibrahimi, S., and ter Hoeve, M. Conditional image generation and manipulation for user-specified content. *arXiv:2005.04909*, 2020.

Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X.-Y., Wu, F., and Bao, B. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv:2008.05865*, 2020.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. *arXiv:1711.00937*, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv:1706.03762*, 2017.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *arXiv:1711.10485*, 2017.

Ye, H., Yang, X., Takac, M., Sunderraman, R., and Ji, S. Improving text-to-image synthesis using contrastive learning. *arXiv:2107.02423*, 2021.

Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., and Yang, Y. Cross-modal contrastive learning for text-to-image generation. *arXiv:2101.04702*, 2021.

Zhang, L., Chen, Q., Hu, B., and Jiang, S. Text-guided neural image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 1302–1310, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10. 1145/3394171.3414017. URL https://doi.org/10.1145/3394171.3414017.

Zhou, S., Gordon, M. L., Krishna, R., Narcomey, A., Fei-Fei, L., and Bernstein, M. S. Hype: A benchmark for human eye perceptual evaluation of generative models, 2019.

Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., and Sun, T. Lafite: Towards language-free training for text-to-image generation. *arXiv:2111.13792*, 2021.

Zhu, M., Pan, P., Chen, W., and Yang, Y. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. *arXiv:1904.01310*, 2019.

A. Evaluation Setup

A.1. Human Evaluations

When gathering human evaluations, we always collect 1,000 pairwise comparisons when evaluating photorealism. We also collect 1,000 comparisons for evaluating caption similarity, except for sweeps over guidance scales where we only collect 500. We use Scale Rapid to collect human comparisons. Through this platform, quality control is performed by periodically showing human evaluators *evaluation tasks*: pre-labeled examples that we believed all labelers should agree on.

When computing wins and Elo scores, we count a tie as half of a win for each model. By doing this, ties effectively dilute the wins of each model.

To compute Elo scores, we construct a matrix A such that entry A_{ij} is the number of times model i beats model j. We initialize Elo scores for all N models as $\sigma_i = 0, i \in [1, N]$. We compute Elo scores by minimizing the objective:

$$L_{\text{elo}} := - \sum_{i,j} A_{ij} \cdot \log \left(\frac{1}{1 + 10^{(\sigma_i - \sigma_j)/400}} \right)$$

A.2. Automated Evaluations

We compute MS-COCO FIDs and other evaluation metrics using 30,000 samples from validation prompts. We use the entire validation set as a reference batch unless otherwise stated, and center-crop the validation images. This cropping matches Ramesh et al. (2021) but is a departure from most previous literature on text-conditional image synthesis, which squeezes images rather than center-cropping them. However, center-cropping is standard practice in the majority of work on unconditional and class-conditional image synthesis, and we hope that it will become standard practice in the future for text-conditional image synthesis as well.

For CLIP score, we employ the CLIP ViT-B/16 model released by Radford et al. (2021), and scale scores by the CLIP logit scale (100 in this case).

B. Hyperparameters

B.1. Training Hyperparameters

For the diffusion stack, we use the same model architecture as the ImageNet 64×64 model from Dhariwal & Nichol (2021), but scale the model width to 512 channels, resulting in roughly 2.3 billion parameters for the visual part of the model. For the text encoding Transformer, we use 24 residual blocks of width 2048, resulting in roughly 1.2 billion parameters. Additionally, we train a 1.5 billion parameter upsampling diffusion model to go from 64×64 to 256×256 resolution. This model is conditioned on text in the same

way as the base model, but uses a smaller text encoder with width 1024 instead of 2048. Otherwise, the architecture matches the ImageNet upsampler from Dhariwal & Nichol (2021), except that we increase the number of base channels to 384. We train the base model for 2.5M iterations at batch size 2048. We train the upsampling model for 1.6M iterations at batch size 512. We find that these models train stably with 16-bit precision and traditional loss scaling (Micikevicius et al., 2017).

Our noised CLIP models process 64×64 images using a ViT (Dosovitskiy et al., 2020) with patch size 4×4. We trained our CLIP models for 390K iterations with batch size 32K on a 50%-50% mixture of the datasets used by Radford et al. (2021) and Ramesh et al. (2021). For our final CLIP model, we trained a ViT-L with weight decay 0.0125. After training, we fine-tuned the final ViT-L for 30K iterations on an even broader dataset of internet images.

We pre-trained GLIDE (filtered) for 1.1M iterations before fine-tuning for another 500K iterations for classifier-free guidance and inpainting. Additionally, we trained a small filtered upsampler model with 192 base channels and 512 text encoder channels for 400K iterations.

B.2. Sampling Hyperparameters

For samples shown in this paper, we sample the base model using 150 diffusion steps, except for inpainting samples where we only use 100 steps. For evaluations, we sample the base model using 250 diffusion steps, since this gives a slight boost in FID.

For the upsampler, we use a special strided sampling schedule to achieve good sample quality with only 27 diffusion steps. In particular, we split the sampling process into five segments, and sample from the following number of evenly-spaced steps within in each segment: $10, 10, 3, 2, 2$. This means that we only sample two timesteps in the range $(800, 1000]$, but 10 timesteps in the range $(0, 200]$. This schedule was found by sweeping over FID on our internal validation set.

C. Comparison to Unnoised CLIP Guidance

Existing work has used the publicly-available CLIP models to guide diffusion models. To get recognizable samples from this approach, it is typically necessary to engineer a set of augmentations and auxiliary losses for the generative process. We hypothesize that this is largely due to the CLIP model's training: it was not trained to recognize the noised or blurry images that are produced during the diffusion sampling process.

To test this hypothesis, we compare a popular CLIP-guided diffusion program (Crowson, 2021a) to our approach based
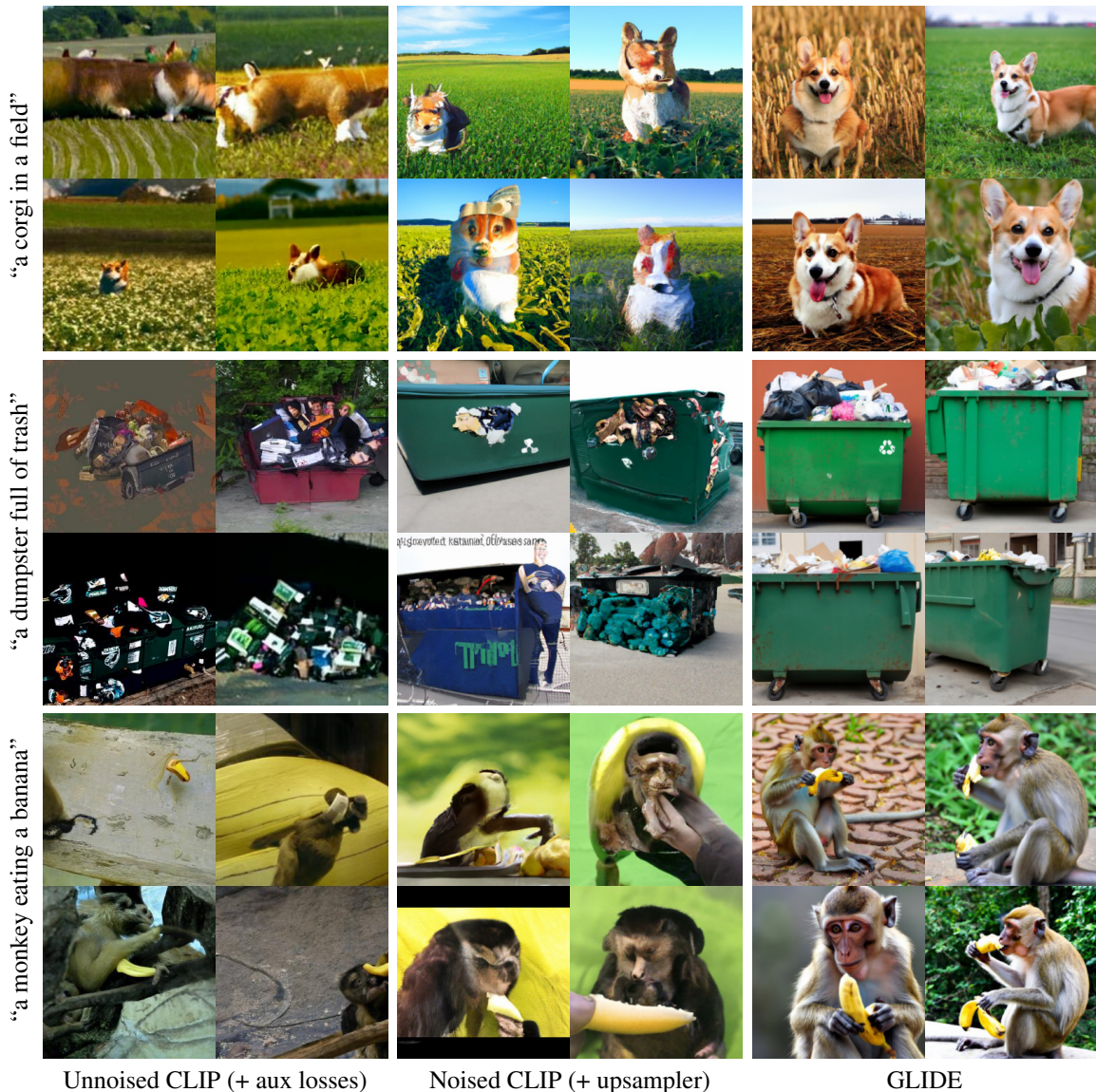
Figure 9. Comparison of GLIDE to two CLIP guidance strategies applied to pre-trained ImageNet diffusion models. On the left, we use a vanilla CLIP model to guide the 256×256 diffusion model from Dhariwal & Nichol (2021), using a combination of engineered perceptual losses and data augmentations (Crowson, 2021a). In the middle, we use our noised ViT-B CLIP model to guide the ImageNet 64×64 diffusion model from Dhariwal & Nichol (2021), then apply a diffusion upsampler. On the right, we show random samples from GLIDE with classifier-free guidance scale 3.0.

on a noised CLIP model (Figure 9). We train a noised ViT-B CLIP model on 64×64 images using the same dataset as Radford et al. (2021). We then use this noised CLIP model to guide a pre-trained ImageNet model towards the text prompt, using a fixed gradient scale of 15.0. Since the ImageNet model is class-conditional, we select a different random class label at each timestep. We then upsample the resulting 64×64 image to 256×256 using our diffusion upsampler. We find that this approach, while much simpler than the approach used by the notebook, produces images of equal or higher quality, suggesting that making CLIP noise-aware is indeed helpful.

D. Inpainting comparison

We evaluate our model and the baselines on prompts shown in (Avrahami et al., 2021) (Figure 10). We find that our fine-tuned model sometimes chooses to ignore the given text prompt and instead produces an image that seems influenced only by the surrounding context. To mitigate this phenomenon, we also evaluate our model with the context fully masked out. This is the inpainting technique first proposed by Sohl-Dickstein et al. (2015), wherein the model only receives information about the context via the noised masked x_t. With this approach, the model seems to fol-
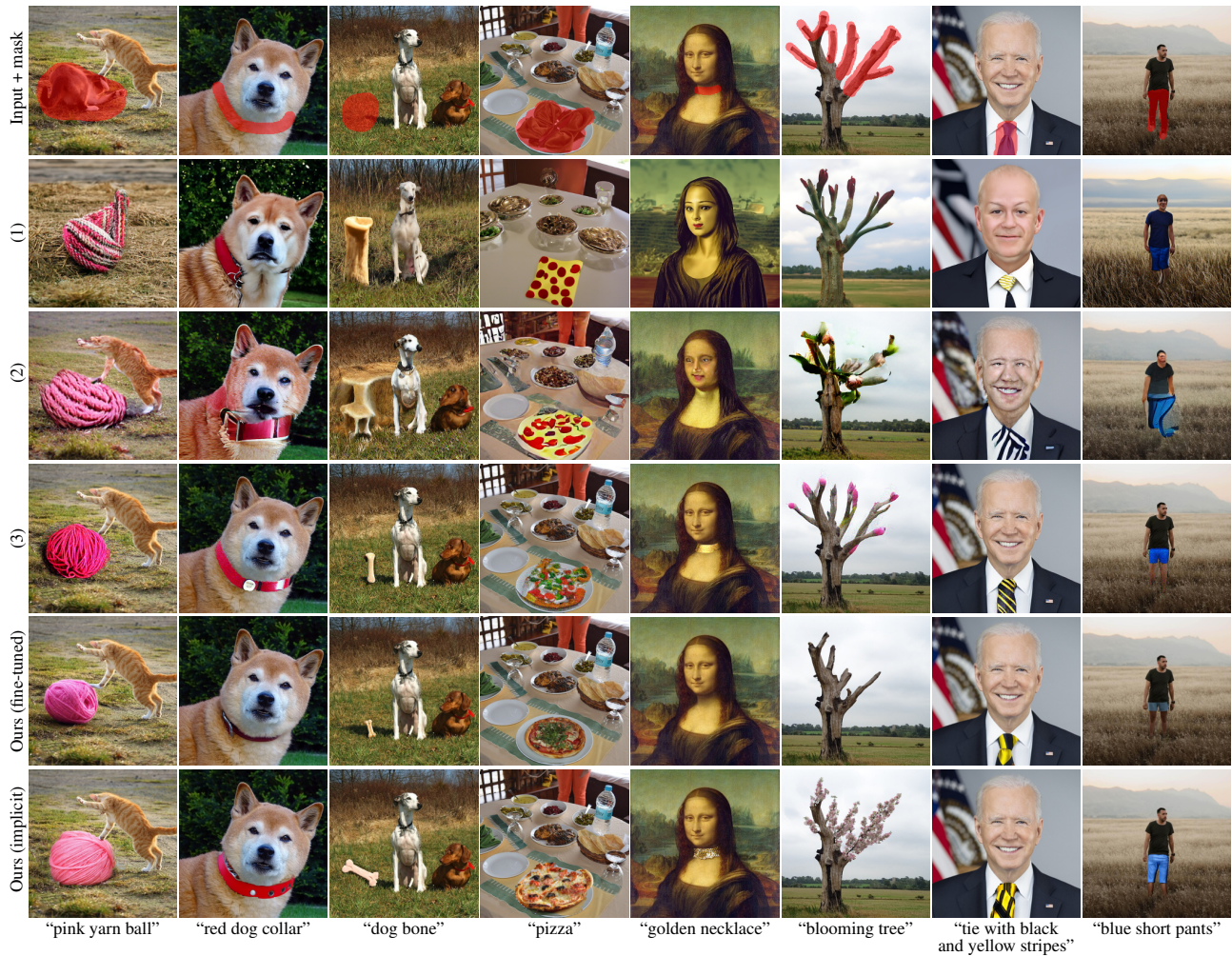
Figure 10. Comparison of image inpainting quality on real images. (1) Local CLIP-guided diffusion (Crowson, 2021a), (2) PaintByWord++ (Bau et al., 2021; Avrahami et al., 2021), (3) Blended Diffusion (Avrahami et al., 2021). For our results, we follow Avrahami et al. (2021) and use CLIP to select the best of 64 samples. Our fine-tuned samples have more realistic lighting, shadows and textures, but sometimes don't focus on the prompt (eg. golden necklace), whereas implicit samples capture the prompt better.

low the caption more consistently, but sometimes produces objects which don't fit as smoothly into the scene.

E. Safety Considerations

Our model is capable of producing fake but realistic images and enables unskilled users to quickly make convincing edits to existing images. As a result, releasing our model without safeguards would significantly reduce the skills required to create convincing disinformation or Deepfakes. Additionally, since the model's samples reflect various biases, including those from the dataset, applying it could unintentionally perpetuate harmful societal biases.

In order to mitigate potentially harmful impacts of releasing these models, we filtered training images before training models for release. First, we gathered a dataset of several hundred million images from the internet, which is largely disjoint from the datasets used to train CLIP and DALL-E, and then applied several filters to this data. We filtered out training images containing people to reduce the capabilities of the model in many people-centric problematic use cases. We also had concerns about our models being used to produce violent images and hate symbols, so we filtered out several of these as well. For more details on our data filtering process, see Appendix F.1.

We trained a small 300 million parameter model, which we refer to as GLIDE (filtered), on our filtered dataset. We then investigated how GLIDE (filtered) mitigates the risk of misuse if the model weights were open sourced. During this investigation, which involved red teaming the model using a set of adversarial prompts, we did not find any instances where the model was able to generate recognizable images of humans, suggesting that our data filter had a sufficiently low false-negative rate. We also probed GLIDE (filtered) for

some forms of bias and found that it retains, and may even amplify, biases in the dataset. For example, when asked to generate "toys for girls", our model produces more pink toys and stuffed animals than it does for the prompt "toys for boys". Separately, we also found that, when prompted for generic cultural imagery such as "a religious place", our model often reinforces Western stereotypes. We also observed that the model's biases are amplified when using classifier-free guidance. Finally, while we have hindered the model's capabilities to generate images in specific classes, it retains inpainting capabilities, the misuse potential of which are an important area for further interdisciplinary research. For detailed examples and images, see Appendix F.2.

The above investigation studies GLIDE (filtered) on its own, but no model lives in a vacuum. For example, it is often possible to combine multiple models to obtain a new set of capabilities. To explore this issue, we swapped GLIDE (filtered) into a publicly available CLIP-guided diffusion program (Crowson, 2021a) and studied the generation capabilities of the resulting pair of models. We generally found that, while the CLIP model (which was trained on unfiltered data) allowed our model to produce some recognizable facial expressions or hateful imagery, the same CLIP model produced roughly the same quality of images when paired with a publicly available ImageNet diffusion model. For more details, see Appendix F.2.

To enable further research on CLIP-guided diffusion, we also train and release a noised ViT-B CLIP model trained on a filtered dataset. We combine the dataset used to train GLIDE (filtered) with a filtered version of the original CLIP dataset. To red team this model, we used it to guide both GLIDE (filtered) and a public 64×64 ImageNet model. On the prompts that we tried, we found that the new CLIP model did not significantly increase the quality of violent images or images of people over the quality of such images produced by existing public CLIP models.

We also tested the ability of GLIDE (filtered) to directly regurgitate training images. For this experiment, we sampled images for 30K prompts in the training set, and computed the distance between each generated image and the original training image in CLIP latent space. We then inspected the pairs with the smallest distances. The model did not faithfully reproduce the training images in any of the pairs we inspected.

F. GLIDE (filtered)

F.1. Data Filtering for GLIDE (filtered)

To remove images of humans and human-like objects from our dataset, we first collect several thousand boolean labels for random samples in the training set. To train the classifier, we resize each image so that the smaller side is 224 pixels,

and then take three crops at the endpoints and middle along the longer side. We feed all three crops into a pre-trained CLIP ViT-B/16, and mean-pool the resulting feature vectors. Finally, we fit an SVM with an RBF kernel to the resulting feature vectors, and tune the bias to result in less than a 1% false negative rate. We tested this model on a separate batch of 1024 samples, and found that it produced no false negatives (i.e. we manually visually inspected the images the model classified as not containing people, and we ourselves found no images of people).

While developing the people filter, we were aiming to detect all people in all types of environments reliably, a task which is often difficult for modern face detection systems especially when dealing with people of all demographics (Buolamwini & Gebru, 2018; Santurkar et al., 2019). In our initial experiments, where we used a ViT-B/32 instead of a ViT-B/16, we observed some cases where people in low-light or obstructed conditions would be missed by the classifier. However, after switching to a ViT-B/16 for feature extraction (which has higher hidden-state resolution than the ViT-B/32), we found that this effect was remedied in all the previously observed failure cases.

To remove images of violent objects, we first used CLIP to search our dataset for words and phrases like "weapon", "violence", etc. After collecting a few hundred positive and negative examples, we trained an SVM similar to the one above. We then labeled samples near the decision boundary of this SVM to obtain another few hundred negative and positive examples. We iterated on this process several times, and then tuned the bias of the final SVM to result in less than a 1% false negative rate. When tested on a separate batch of 1024 samples, this classifier produced no false negatives.

We initially approached the removal of hate symbols the same way, using CLIP to search the dataset for particular keywords. However, we found that this approach surfaced very few relevant images, suggesting that our data sources had already filtered for this content in some way. Nonetheless, we used a search engine to collect images of two prevalent hate symbols in America, the swastika and the confederate flag, and trained an SVM on this data. We used the active learning procedure described above to collect more negative examples near the decision boundary (but could not find positive ones), and tuned the resulting SVM's bias to result in less than a 1% false negative rate on this curated dataset.

F.2. Biases and CLIP Guidance for GLIDE (filtered)

GLIDE (filtered) continues to exhibit bias – a demonstration both of how biases in image datasets extend beyond those found in images of people, and pointing to biases in the choices we made in filtering. For example, the model produces different outputs when asked to generate toys for

(a) "toys for boys"

(b) "toys for girls"

Figure 11. GLIDE (filtered) samples for the same random seed when changing the gender in the prompt.

(a) Unguided

(b) Classifier-free guidance (scale 3.0)

Figure 12. GLIDE (filtered) samples for the prompt "a religious place" using the same random seed, but with different guidance scales.

boys and toys for girls (Figure 11). When asked to generate "a religious place", the model tends to gravitate towards church-like buildings, and this bias is amplified by classifier-free guidance (Figure 12).

We expect that our hate symbol classifier has a strong American and Western bias, since it was only trained on two prevalent hate symbols in America. As a result, it is likely that the training data retains images depicting other hateful symbols. However, we do find that the model generally has diminished capabilities across a wider set of symbols (Figure 13). We hypothesize that this may be a result of the smaller dataset available to GLIDE (filtered).

We also incorporated GLIDE (filtered) into a publicly-available CLIP-guided diffusion program (Crowson, 2021a). We found that the resulting combination had some ability to generate face-like objects (e.g. Figure 14). While, the original CLIP-guided diffusion program using a publicly-available diffusion model often produced more recognizable images in response to our prompts, these findings highlight one of the limitations of our filtering approach. We also found that GLIDE (filtered) still exhibits a strong Western bias in some cases, often exceeding the bias exhibited by the existing publicly-available diffusion model (Figure 15).
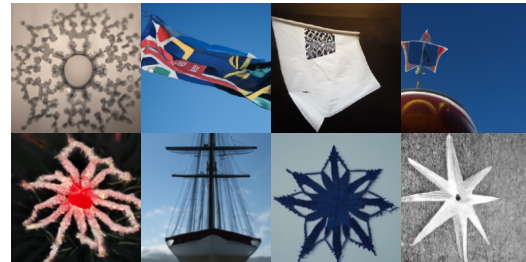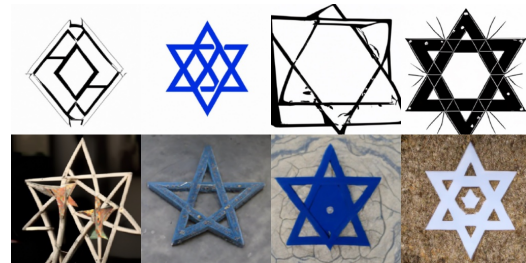

(a) GLIDE (filtered)

(b) GLIDE (small)

Figure 13. Comparison of GLIDE (filtered) and GLIDE (small) samples for the prompt "the star of david".

(a) GLIDE (filtered)

(b) GLIDE (filtered) + unfiltered public CLIP

(c) Baseline (publicly-available models)

Figure 14. Generations for the prompt "a happy person". In (a), we show GLIDE (filtered) with classifier-free guidance scale 3.0. In (b), we use a publicly-available CLIP model to guide GLIDE (filtered). In (c), we use a publicly-available CLIP model to guide a publicly-available ImageNet diffusion model.

(a) GLIDE (filtered)

(b) GLIDE (filtered) + unfiltered public CLIP

(c) Baseline (publicly-available models)

Figure 15. Generations for the prompt "a place of worship". In (a), we show GLIDE (filtered) with classifier-free guidance scale 3.0. In (b), we use a publicly-available CLIP model to guide GLIDE (filtered). In (c), we use a publicly-available CLIP model to guide a publicly-available ImageNet diffusion model.

G. Comparison to Smaller Models

Was it worth training our large GLIDE model? To answer this question, we train another 300 million parameter model (referred to as GLIDE (small)) on our full dataset using the same hyperparameters as GLIDE (filtered). We compare samples from our large, small, and safe models to determine what capabilities we gain from training such a large model on a large, diverse dataset.

In Figure 16, we observe that the smaller models often fail at binding attributes to objects (e.g. the corgi) and perform worse at compositional tasks (e.g. the blocks). All of the models can often produce realistic images, but the two models trained on our full dataset are much better at combining unusual concepts (e.g. a hedgehog using a calculator).

We also conduct a human evaluation comparing our small and large models with and without classifier-free guidance. We first swept over guidance scales for the 300M model using a human evaluation, finding that humans slightly prefer scale 4.0 to 3.0 for this small model. We then ran a human evaluation comparing both models with and without guidance (Table 4). We find that classifier-free guidance gives a larger Elo boost than scaling the model by roughly 10x.

Table 4. Elo scores resulting from a human evaluation comparing our small model to our larger model

Size	Guide. Scale	Photorealism	Caption
300M	1.0	-131.8	-136.4
300M	4.0	28.2	70.9
3.5B	1.0	-23.9	-27.1
3.5B	3.0	**133.0**	**140.5**

H. Additional Samples

In Figures 17 and 18 we show 4×4 grids of random samples from our model with no guidance, classifier-free guidance, and CLIP guidance using the same random seeds, as well as samples from DALL-E. We find that classifier-free guidance produces the highest-quality images most reliably. For DALL-E, we sample 512 images for each prompt and select the top 16 using CLIP reranking. For all other sample grids, we show 16 random samples without CLIP reranking.
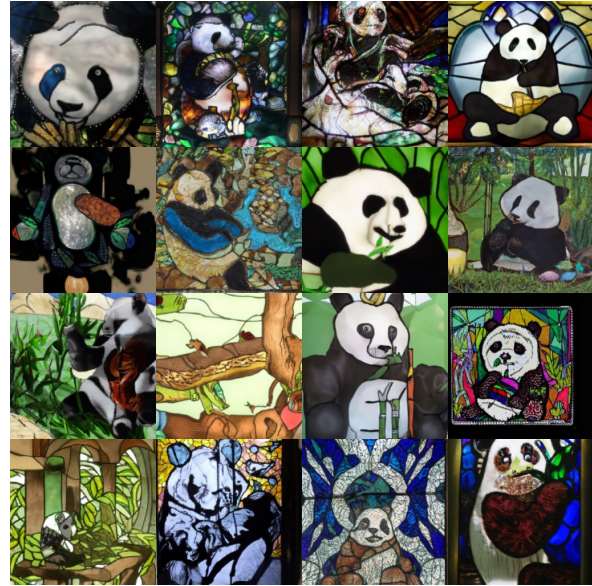
"a hedgehog using a calculator" "a corgi wearing a red bowtie and purple party hat" "a red cube on top of a blue cube" "a high-quality oil painting of a psychedelic hamster dragon"

Figure 16. Comparing classifier-free guided samples from our large model (first row), a small version trained on the same data (second row), and our released small model trained on a smaller, filtered dataset. In the final row, we show samples using our small model guided by a CLIP model trained on filtered data. Samples are not cherry-picked.

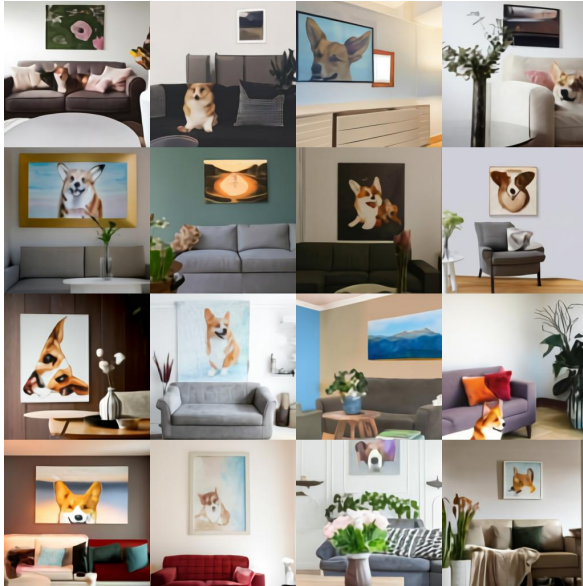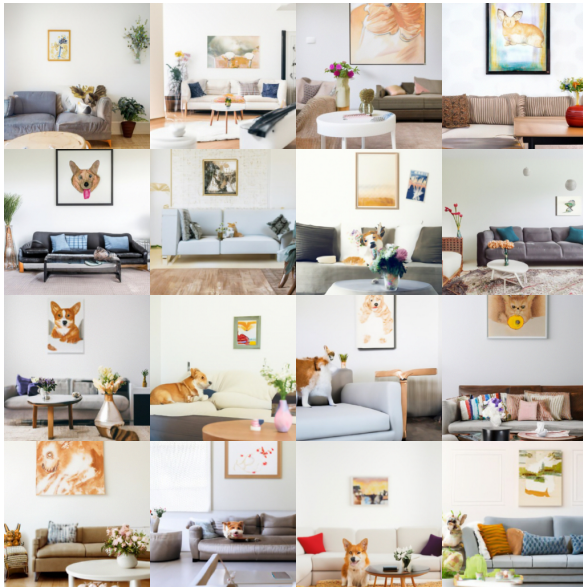(a) DALL-E (Temp 0.85, CLIP reranked top 16 out of 512)

(b) GLIDE (Unguided)

(c) GLIDE (CLIP guidance, scale 2.0)

(d) GLIDE (Classifier-free guidance, scale 3.0)

Figure 17. Random samples from DALL-E and GLIDE on the prompt "a stained glass window of a panda eating bamboo". We do not perform any CLIP reranking for GLIDE.
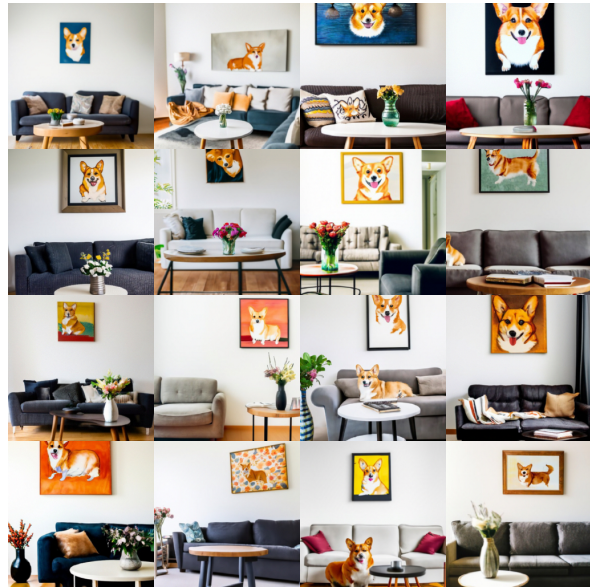
(a) DALL-E (Temp 0.85, CLIP reranked top 16 out of 512)

(b) GLIDE (Unguided)

(c) GLIDE (CLIP guidance, scale 2.0)

(d) GLIDE (Classifier-free guidance, scale 3.0)

Figure 18. Random samples from DALL-E and GLIDE on the prompt "A cozy living room with a painting of a corgi on the wall above a couch and a round coffee table in front of a couch and a vase of flowers on a coffee table". We do not perform any CLIP reranking for GLIDE.