

TeTIm-Eval: a novel curated evaluation data set for comparing text-to-image models

Federico A. Galatolo¹^a, Mario G.C.A. Cimino¹^b and Edoardo Cogotti¹

¹*Department of Information Engineering, University of Pisa, 56122 Pisa, Italy*
federico.galatolo@ing.unipi.it, mario.cimino@unipi.it

Keywords: Text-to-Image Model, Deep Learning, Curated Data Set, Generative Image Model

Abstract: Evaluating and comparing text-to-image models is a challenging problem. Significant advances in the field have recently been made, piquing interest of various industrial sectors. As a consequence, a gold standard in the field should cover a variety of tasks and application contexts. In this paper a novel evaluation approach is experimented, on the basis of: (i) a curated data set, made by high-quality royalty-free image-text pairs, divided into ten categories; (ii) a quantitative metric, the CLIP-score, (iii) a human evaluation task to distinguish, for a given text, the real and the generated images. The proposed method has been applied to the most recent models, i.e., DALLE2, Latent Diffusion, Stable Diffusion, GLIDE and Craiyon. Early experimental results show that the accuracy of the human judgement is fully coherent with the CLIP-score. The dataset has been made available to the public.

1 INTRODUCTION

In recent years, the field of generating images from text has seen unprecedented growth, with numerous text-to-image techniques being developed (Frolov et al., 2021). Even if conditional Generative Adversarial Neural Networks were one of the first deep learning-based architectures proposed for the text-to-image task (Zhang et al., 2018) (Reed et al., 2016) their promising results are generally limited to low-variability data, as their adversarial learning approach does not scale well to modeling complex, multi-modal distributions (Brock et al., 2018). Diffusion Models (DMs), which are constructed from a hierarchy of denoising autoencoders, have recently achieved impressive results in image synthesis and beyond, defining the state-of-the-art in class-conditional image synthesis and super-resolution (Dhariwal and Nichol, 2021) (Saharia et al., 2021). Furthermore, unconditional DMs can be used for tasks such as inpainting and colorization, as well as stroke-based synthesis (Song et al., 2020). Simple autoregressive transformers, such as the first version of DALLE, on the other hand, have also produced good results in this field (Ramesh et al., 2021).


In addition, the Contrastive Language-Image Pre-


Training (CLIP) neural network has recently emerged (Radford et al., 2021). CLIP has been trained on a wide range of image and text pairs. It can be instructed in natural language to predict the most relevant text snippet given an image without directly optimizing for the task. CLIP embeddings are also robust to picture distribution change, and have been used to achieve state-of-the-art results on vision and language tasks (Shen et al., 2021) as well as image generation from text, when combined with Diffusion Models.

In this paper we are going to compare the performance of the most recent text-to-image architectures. In particular we are going to consider DALLE-2, Latent Diffusion, Stable Diffusion, GLIDE and craiyon (formerly known as DALL-E mini).

To evaluate the models, we created a new high-quality dataset called TeTIm-Eval (Text To Image Evaluation) composed of 2500 labelled images and 300 text-image pairs divided into ten classes. The selected models were then used to generate images from the dataset captions, and quantitative evaluation metrics such as the Fréchet Inception Distance (FID) and the CLIP-score were computed with respect to the ground truth images. Finally, we presented the caption and image from the dataset, as well as a generated image, to human evaluators to assess each model's ability to generate realistic images.

The remainder of the paper is structured as follows: Section 2 describes the considered models,

^a <https://orcid.org/0000-0001-7193-3754>

^b <https://orcid.org/0000-0002-1031-1959>

the proposed dataset, the quantitative evaluation metrics, the design of the human experimentation, and the methods for processing the results. Section 3 illustrates the experimental results, and Section 4 discusses the findings and future research.

2 MATERIALS AND METHODS

2.1 Generative models

In this paper we considered five text to image models: DALLE-2, Latent Diffusion, Stable Diffusion, GLIDE and crayon. The first four are DM based whereas the latter is a decoder-only autoregressive transformer. We excluded the Imagegen model from our study because, while the model description is publicly available (Saharia et al., 2022), access to the model is not; we also excluded the midjourney model because, while it is accessible online, its implementation is not publicly available.

2.1.1 Diffusion Models

Diffusion models have been inspired by non-equilibrium thermodynamics. By first defining a Markov chain of diffusion steps to gradually introduce random noise to data, they are trained to reverse the diffusion process in order to create desired data samples from the noise. Several diffusion-based generative models, such as diffusion probabilistic models (Sohl-Dickstein et al., 2015) and denoising diffusion probabilistic models (Ho et al., 2020), have been proposed.

Let us define a forward diffusion process for a data point sampled from a real data distribution $x_0 \sim q(x_0)$. In this process we gradually add small amounts of Gaussian noise to the sample as shown in Equation 1 and 2, resulting in a series of increasingly noisy samples $x_0, x_1 \dots x_T$. The step sizes are governed by a variance schedule $\beta_t \in (0, 1)$ which can be learned using the reparameterization trick or held constant as hyperparameter.

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

$$q(x_{1:T} | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}) \quad (2)$$

The training objective of the diffusion models is, then, to learn the reverse diffusion process $p_\theta(x)$, which is modeled as a Markov chain as shown in Equation 3 4. This Markov chain that starts with random noise $p_\theta(x_T) = \mathcal{N}(x_T; 0; I)$ and progresses

through a succession of less and less noisy samples $x_T, x_{T-1} \dots x_0$.

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1} | x_t) \quad (3)$$

$$p_\theta(x_{t-1} | x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (4)$$

2.1.2 GPT models for text-to-image

Generative Pre-trained Transformers are a decoder-only transformer based architecture (Vaswani et al., 2017) which uses an unidirectional self-attentive model that attends just the tokens before a each token in a sequence (Radford et al., 2018). GPT models' training objective is then to predict the next token given the previous ones as context.

Despite the fact that GPT models are typically utilized to handle natural language processing tasks, they can be trained to model text and image tokens as a single stream of data in a autoregressive fashion. Using pixels directly as image tokens, on the other hand, would need an excessive amount of memory for high-resolution photos. GPT-based text to image models solved this issue with using two-stage models.

The first stage is to train a discrete variational autoencoder (dVAE) to compress each image into a much smaller grid of image tokens. In the second stage, a GPT transformer is trained on sequences created by concatenating text and image tokens to model the joint distribution over text and images. Finally, to complete the text to image task, the model is given the target text tokens and is used to predict the subsequent image tokens. The predicted image tokens are then fed into the dVAE decoder and projected into the RGB space (Ramesh et al., 2021).

2.1.3 CLIP

CLIP is a neural network that was trained on a large set of image and text pairs (400M). CLIP can be used to find the text snippet that best represents a given image, or the most appropriate image given a text query, as a result of this multi-modality training.

In CLIP an image encoder and a text encoder were trained simultaneously to predict the correct pairing of a set of images and text. They were trained to predict, given an image, which one of the randomly sampled text snippets the image was paired to in the training dataset and vice-versa. The model should abstract multiple concepts from the images and from the texts in order to solve the task and the resulting encoders should produce similar embeddings if the image and the text contains similar visual and textual

concepts. This approach differs significantly from traditional image tasks, in which the model is typically required to identify a class from a large set of classes (e.g. ImageNet).

More formally give a set of images x_0, x_1, \dots, x_{n-1} and the respective captions y_0, y_1, \dots, y_{n-1} the features vectors I_0, I_1, \dots, I_{n-1} are computed using the image encoder $I_i = IE(x_i)$ and the features vector T_0, T_1, \dots, T_{n-1} are computed using the text encoder $T_i = TE(y_i)$. Finally the logits cross product is computed as $l = (I \otimes T) \cdot e^\tau$ and the encoder and decoder are trained to solve n joint classification problems (e.g. classify each feature vector I_i with the respective feature vector T_i)

2.1.4 Models taken into account

In this paper we considered five text to image models: DALLE-2, Latent Diffusion, Stable Diffusion, GLIDE and Craiyon.

DALLE-2 architecture (Ramesh et al., 2022) is composed by two stages: a prior and a decoder. Given a training dataset of pairs (x, y) of images x and their corresponding captions y , and indicating the CLIP image and text embeddings with z_i and z_t , respectively. The prior $P(z_i|y)$ is trained to generate CLIP image embeddings z_i based on captions y , and the decoder $D(x|z_i)$ is trained to generate images x based on CLIP image embeddings z_i . The prior then learns a generative model of the image embeddings, whereas the decoder inverts images based on their CLIP image embeddings.

For the prior network the continuous vector z_i is directly modelled using a gaussian diffusion model conditioned on the caption y . The decoder $D(x|z_i)$ is also modeled with a gaussian diffusion model as in (Nichol et al., 2021), but with four additional context tokens encoding the CLIP embeddings concatenated to the text encoder output. Finally, to obtain high resolution images, the output of the decoder is passed to a two-stage diffusion upsampler model which upsamples the output from 64x64 to 256x256 and from 256x256 to 1024x1024.

Latent Diffusion and Stable Diffusion (Rombach et al., 2022) are the same architecture trained on different dataset using a different set of hyperparameters. Latent Diffusion is a two-stage architecture. The first stage is a self-supervised encoder-decoder architecture where the encoder $z = E(x)$ encodes the image x into the embeddings z and the decoder $\hat{x} = D(z)$ decodes the embeddings z back into the original image \hat{x} . The second stage is a diffusion process on the latent space z conditioned on image captions $z_T = LDM(z_{T-1}|y)$. The output image is then computed using the decoder on the last step of the diffusion pro-

cess $D(Z_t)$.

A Vector Quantized Variational AutoEncoder (VQ-VAE) (Van Den Oord et al., 2017) (Esser et al., 2021) trained by combining a perceptual loss and a patch-based adversarial objective was used for the first stage. This guarantees that the reconstructions are restricted to the images manifold by imposing local realism and avoids the blurriness induced by relying exclusively on pixel-space losses. A Vector Quantized regularization was also used to avoid high variance latent space. It was used an encoder with a downsample factor of 8.

The second stage is a denoising UNet (Ronneberger et al., 2015) conditioned on text embeddings implementing a diffusion process in the latent space z . The CLIP ViT-L/14 text encoder is used to compute the text embeddings, and the conditioning is achieved by augmenting the UNet backbone with the cross-attention mechanism. y is encoded with a modality-specific encoder and used to compute keys and values for the attention layers to support conditioning from various modalities (text, semantic maps, etc.); as queries, the flattened internal states of the UNet have been used.

Guided Language to Image Diffusion for Generation and Editing (GLIDE) (Nichol et al., 2021) is a large diffusion model from OpenAI. GLIDE implements a text-conditioned diffusion model directly on the pixel space. It implements an Ablated Diffusion Model (ADM) as proposed in (Dhariwal and Nichol, 2021) augmented with text conditioning information. Formally for each image of the forward diffusion process x_t and corresponding text caption y , GLIDE predicts $DM(x_{t-1}|x_t, y)$. To condition the diffusion model with text the descriptive caption y is encoded using a Transformer and the last embedding of the sequence is used in place of the class conditional token in the ADM architecture; moreover the overall last layer embeddings are projected to the correct dimension and concatenated to the attention context of each ADM layer. Finally each generated images is upsampled from a 64x64 dimension to a 256x256 using a upsample diffusion model.

Craiyon (Craiyon,) (formerly known as DALLE mini) is community trained reduced instance of a DALLE (Ramesh et al., 2021) model. Craiyon uses a two-stage training method. As in Latent and Stable Diffusion, the first stage is a self-supervised encoder-decoder architecture. Craiyon uses a VQGAN (Esser et al., 2021) with a reduction factor of 16 pretrained on the ImageNet dataset for this stage. A GPT-like autoregressive transformer is used in the second stage. A BPE tokenizer is used to encode the text caption y , and the target image is encoded in 32x32=1024 image

tokens. The text tokens (up to 256) and image tokens are then concatenated and used to train the autoregressive transformer. The overall procedure is equivalent to maximizing the evidence lower bound on the model distribution’s joint likelihood over images x and captions y .

2.2 TeTIm-Eval dataset

To assess the performance of the models under evaluation, we created a diverse and high-quality dataset of text and image pairs that we called TeTIm-Eval (Text To Image Evaluation). Existing datasets and category-specific websites were used to create the dataset. We considered three types of images: paintings, drawings, and realistic photographs. And, as shown in Figure 1, we identified the following sub-categories for each category.

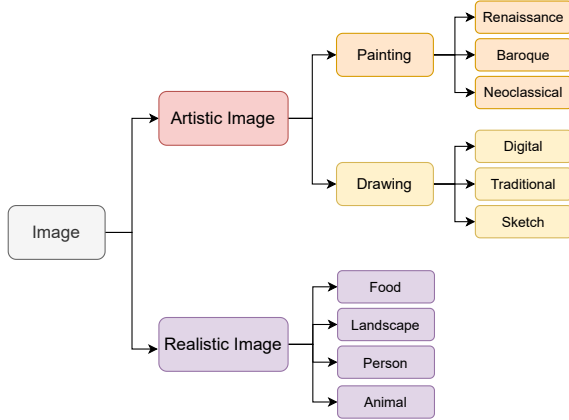


Figure 1: TeTIm-Eval categories and sub-categories taxonomy

- Painting
 - Renaissance paintings
 - Baroque paintings
 - Neoclassical paintings
- Drawing
 - Digital Drawings
 - Traditional Drawings
 - Sketch Drawings
- Realistic Photos
 - Food Photos
 - Landscape Photos
 - Person Photos
 - Animal Photos

We first identified the sources and downloaded the available data, then sampled the downloaded data at

random and manually rejected images that did not meet our quality criteria, yielding a total of 2500 images (250 per sub-category). Finally, in order to create the final dataset, we randomly selected 30 images from each sub-category and manually wrote a textual description for each of the 300 images.

We followed these quality criteria to ensure the creation of an high-quality dataset:

- Only images that clearly belong to the sub-category
- Only images where the content fills the available space (no frames, etc.)
- Only images released under the Creative Commons License
- Only high definition images
- No signatures or watermarks
- No adult-rated, violent or hate images
- No political images

Category	Sub-category	Source
Painting	Renaissance	Wikiart
	Baroque	Wikiart
	Neoclassical	Wikiart
Drawing	Digital	Deviantart and Openverse
	Traditional	Deviantart and Openverse
	Sketch	ImageNet Sketch
Realistic	Food	Wikimedia Commons
	Landscape	Wikimedia Commons
	Person	COCO
	Animal	Wikimedia Commons

Table 1: Dataset sources

We selected a total of 6 different data sources: Wikiart (Wikiart,) which is an online, user-editable visual art encyclopedia, Deviantart (deviantart,) which is an online art community, Openverse (openverse,) which is an open-source search engine for open content developed as part of the WordPress project, ImageNet Sketch (Wang et al., 2019) which is a dataset consisting of 50000 images (50 images for each of the 1000 ImageNet(Deng et al., 2009) classes), Wikimedia Commons (wikimedia,) which is a media repository of open images, sounds, videos and other media from the Wikimedia Foundation and COCO (Lin et al., 2014) which is a large-scale object detection, segmentation, and captioning dataset. All the painting images were sampled from Wikiart, digital and traditional drawings from Deviantart and Openverse, the sketches from ImageNet Sketch, the food landscape and animal photos from Wikimedia Commons and finally the person photos from the

COCO dataset as shown in Table 1. Finally the captions were written by the same person to ensure consistency across the whole dataset. Furthermore given the dataset’s curated and high-quality nature, it can also be used to train and/or evaluate zero/few shot learning algorithms, which are notorious for requiring high-quality data to perform well.

The overall dataset, the 2500 labelled images and the 300 text-image pairs, as well as its companion source code has been released on GitHub (Galatolo, 2022).

3 EXPERIMENTAL RESULTS AND DISCUSSION

We compared the models taken into account using 300 image-text pairs and three evaluation metrics: the CLIP score, the Fréchet Inception Distance (FID), and the human evaluation. Specifically, firstly, using the CLIP text encoder, we computed the feature vector of each target text $TE(y)$. Then, using the CLIP image encoder $IE(x_i)$, we computed the CLIP score as its dot product with the feature vectors of the generated images. Formally given a target text y and a set of generated images x_0, x_1, \dots, x_{n-1} the CLIP score of the image x_i with respect to the target text y is $CS_i = TE(y) \cdot IE(x_i)$. Even if the CLIP score is a good metric to determine the semantic distance between captions and images it is important to point out that some of the models under study directly use CLIP as part of their pipeline like DALLE2 or use CLIP indirectly to filter the training dataset (like Stable and Latent Diffusion)

The Fréchet Inception Distance(FID) is a widely used to assess the quality of images created by a generative models(Heusel et al., 2017). The FID is the Multivariate Gaussian Fréchet distance of the probability distributions of the features extracted using an Inception V3 (Szegedy et al., 2016) model pre-trained on the ImageNet dataset. Formally given the probability distribution of the inception features extracted from the real images $\mathcal{N}(\mu_r, \Sigma_r)$ and the one of the inception features extracted from the generated images $\mathcal{N}(\mu_g, \Sigma_g)$ the FID can be computed as $FID = |\mu_r - \mu_g| + tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}})$. Even if the FID is the most used metric to compare text to image models it has been shown (Chong and Forsyth, 2020) that for a low number of samples it is not reliable, because it frequently fails to reflect the goodness and fidelity of generated images. For this reason it is not used in the current experimentation. For the Human evaluation we developed a web platform in

which each user was prompted with a descriptive caption and two images: the real image from the TeTIm-Eval dataset and an image generated by one of the models under study. We then asked the user to identify the real image.

Title	Value
Participants	183
Answers	5010
Right Answers	3419

Table 2: Overall human evaluation statistics

Metric	Value
Accuracy	.68
False Positive Rate	.34
False Negative Rate	.30
Precision	.67
Recall	.70

Table 3: Overall human evaluation performance

Table 2 and table 3 show the human involved and their answers, as well as their overall performance, respectively. Table 4 shows the human performance per category. Finally Table 5 displays the human performance per model as well as its CLIP-score, showing that human accuracy is consistent with the CLIP score. It is important to note that in this context, accuracy is defined as the number of correctly identified human-generated images divided by the total number of answers. As a result, the lower the accuracy, the better the model is at producing images that fool humans into thinking they were created by humans. In this context, the lower the human accuracy, the better the model.

Category	Sub-Category	H. Acc.	H. Prec.	H. Rec.
Painting	Renaissance	.73	.73	.74
	Baroque	.77	.74	.80
	Neoclassical	.73	.69	.79
Drawing	Digital	.63	.68	.49
	Traditional	.60	.63	.56
	Sketch	.62	.58	.66
Realistic	Food	.66	.62	.77
	Landscape	.65	.61	.70
	Person	.77	.76	.80
	Animal	.68	.66	.74

Table 4: Overall human performance per category

Model	H. Accuracy (\downarrow)	CLIP Score (\uparrow)
Stable Diffusion	.62	29
DALLE2	.63	29
Craiyon	.71	28
Latent Diffusion	.71	26
GLIDE	.73	22

Table 5: Overall human performance per model

4 CONCLUSIONS

In this paper, a novel evaluation method for text-to-image models is introduced. A novel high-quality evaluation dataset consisting of 2500 labelled images and 300 text-image pairs has been released to the public. Five different state-of-the-art models have been evaluated, both using a quantitative metric and humans involving about 180 participants, for 5K+ answers. The human evaluation shows that Stable Diffusion is the most accurate model, followed by DALLE2, Craiyon, Latent Diffusion and GLIDE. The CLIP-score metric is coherent with the human evaluation. Future work will consider a larger data set and will include the FID metrics. Given the high quality of the dataset, it is worth noting that it can also be used to train and/or evaluate zero/few shot learning algorithms.

ACKNOWLEDGEMENTS

This work has been partially supported by: (i) the National Center for Sustainable Mobility MOST/Spoke10, funded by the Italian Ministry of University and Research in the framework of the National Recovery and Resilience Plan; (ii) the PRA_2022_101 project “Decision Support Systems for territorial networks for managing ecosystem services”, funded by the University of Pisa; (iii) the Ministry of University and Research (MUR) as part of the PON 2014-2020 “Research and Innovation” resources – Green/Innovation Action – DM MUR 1061/2022”; the MUR, in the framework of the FISR 2019 Programme, under Grant No. 03602 of the project “SERICA”; the Tuscany Region in the framework of the SecureB2C project, POR FESR 2014-2020, Project number 7429 31.05.2017; the MUR, in the framework of the “Reasoning” project, PRIN 2020 LS Programme, Project number 2493 04-11-2021.

REFERENCES

- Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis.
- Chong, M. J. and Forsyth, D. (2020). Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079.
- Craiyon. Craiyon website. <https://www.craiyon.com/>. Accessed: 2022-12-04.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- deviantart. Deviantart website. <https://www.deviantart.com/>. Accessed: 2022-12-04.
- Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis.
- Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.
- Frolov, S., Hinz, T., Raue, F., Hees, J., and Dengel, A. (2021). Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209.
- Galatolo, F. A. (2022). Tetim-eval repository. <https://github.com/galatolofederico/tetim-eval>.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.
- openverse. Openverse website. <https://wordpress.org/openverse/>. Accessed: 2022-12-04.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.
- Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2021). Image super-resolution via iterative refinement.
- Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. (2021). How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, H., Ge, S., Lipton, Z., and Xing, E. P. (2019). Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518.
- Wikiart. Wikiart website. <https://www.wikiart.org/>. Accessed: 2022-12-04.
- wikimedia. Wikimedia commons website. <https://commons.wikimedia.org>. Accessed: 2022-12-04.
- Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2018). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962.