Analysis of Faces and Facial Details Generated by Text-to-Image Models

Xiaona Zhou
Virginia Tech
Xzhou1@vt.edu

Abstract

In this project, we reproduced the results of Borji [1] and extended the work to further analysis of facial details, eyes and mouth. We also included a new diffusion model, ERNIE-ViLG, that achieves state-of-the-art on MS-COCO with zero-shot FID score of 6.75, which outperforms Google's Imagen. We find that despite being the state-of-the-art Text-to-Image model, ERNIE-ViLG scores highest in FID score on all categories. While DALL-E 2 struggles more with generating natural eyes, Midjourney falls short on mouth details. we explored possible reasons and included some analysis of the model.

1 Introduction

Some qualitative analyses on Text-to-Image models have been focused on their shortcomings, such as generating images that involve printed text and contain more than one person, lacking knowledge about scientific facts, and replicating societal stereotypes, as discussed in these post articles Romero [2], STRICKLAND [3], Swimmer963 [4]. In terms of quantitative analysis, Cho et al. [5] proposed ways of evaluating Text-to-Image generative models against reasoning skills (object detection, object counting, and spatial relation understanding), and social bias (gender and race). Borji [1] was the first to quantitatively evaluate faces generated by these models. This project replicates and extends the work by including the state-of-the-art Text-to-Image model ERNIR-ViLG and analyzing facial details in terms of FID scores. We are interested to see exactly where these models fail to produce a realistic face.

2 Comparison

2.1 Models

Other than the three models, Stable Diffusion, DALL-E 2, and Midjourney, examined by Borji [1], we also included ERNIE-ViLG.

- **ERNIE-ViLG[6].**This is a large-scale Chinese text-to-image diffusion model that achieves state-of-the-art on MS-COCO with a zero-shot FID score of 6.75, which outperforms Google's ImagenSaharia et al. [7]. Novelty of ERNIE-ViLG includes using text and visual knowledge from the image during learning, and "mixture-of-denoising-experts", instead of one U-Net for all steps, for denoising. ERNIE-ViLG 2.0 contains a transformer-based text encoder with 1.3B parameters and 10 denoising U-Net experts with 2.2B parameters each, which totally adds up to about 24B parameters. The model was trained on LAION Schuhmann et al. [8] and some internal Chinese datasets (translated into Chinese with Baidu

Preprint. Under review.

Translate) on 320 A100 GPUs for 18 days. The demo is available on Hugging Face[1]. We follow the instructions on their GitHub page[2] to run ERNIE-ViLG.

2.2 Data

We used face images dataset, GFW, [3] provided by Borji [1] for comparisons on Stable Diffusion, DALL-E 2, and Midjourney. As a practice, we also extracted face images and captions from COCO dataset. We used the extracted caption for image generation with ERNIE-ViLG.

To evaluate the models in respect of FID scores, we need two sets of images, real images, and generated images. We evaluate the models in three categories: a) face, b) eyes, c) mouth.

- **Real images**
  - **face images.** We ran a face detector on the COCO training set (train2017 and train2014) on the category "person" to extract face images and corresponding captions. We used the Multi-Task Cascaded Convolutional Neural Network Zhang et al. [9], or MTCNN, for face detection. The two features of this library were found useful for our task, confidence level, and the dimension of the bounding box. We used two criteria to make sure that the faces extracted are of high quality. The confidence level was set to 0.99 to make sure that the face detected was indeed a face. The bounding box was set to have the width and height differ by at least 15, which ensures that the face extracted is clear enough. During extraction, for face images that satisfy the conditions, we kept one of the five captions and resize it to 250×250. In total, we collected 10,000 real faces.
  - **eyes images.** To obtain eyes images, we ran a face detection again on the face images, and cropped out the area near the eyes. The code for extracting eyes can be found on project GitHub page [4]. We only kept the ones that if the difference between two eyes is at least 100, and the two eyes are at about the same level (We ran a few experiments and picked values that extract eyes well). We collected 2,558 eyes images from COCO, and 2,000 images from GFW.
  - **mouth images.** With a similar method as with eyes images, we collected 2,000 mouth images each from COCO and GFW.
- **Generated images.** To generate images, we use the captions we obtained when extracting real face images from COCO dataset.
  - **face images.** we were not able to massively generate images using ERNIE-ViLG due to the following reasons, a) no offline model available. We can only obtain images from their API. b) Limited access to the API. One must have API Key and Secret Key from Baidu to run their code, and there is 100 images/day/account, a maximum of 500 images per account. Also, a China phone number is required to apply for an account from Baidu. What makes the matter worse is that one cannot buy their service without an ID issued by the Chinese government. Despite all these, we still managed to generate 1506 images and collected 750 face images.
  - **eyes and mouth images.** We applied the same extraction function to extract eyes and mouth images from generated face images. We extracted 308 eyes images and 582 mouth images for ERNIE-ViLG.

    As with GFW, we also manually removed the faces that did not meet the criteria for images generated by ERNIE-ViLG. We did not manually remove faces image that we extracted from COCO due to the amount of time required.

2.3 Evaluation Scheme

We used the FID score Seitzer [10] to quantitatively evaluate the images generated by the models, similar to what was done in Borji [1]. We reproduced the FID scores plot in Borji [1] with an

[1] https://huggingface.co/spaces/PaddlePaddle/ERNIE-ViLG

[2] https://github.com/PaddlePaddle/PaddleHub/tree/develop/modules/image/text$_to_image/ernie_vilg$

[3] https://drive.google.com/file/d/1EhbUK64J3d0$_chmD2mpBuWB-Ic7LeFlP/view$

[4] https://github.com/XiaonaZhou/Analysis-of-Faces-and-Facial-Details-Generated-by-Text-to-Image-Models

Table 1: FID scores of models

Name	face	eyes	mouth
Real Image	6.859 ± 0.079	19.565 ± 0.170	20.006 ± 0.175
Stable Diffusion	31.237 ± 0.359	59.001 ± 1.146	62.641 ± 0.848
DALL-E 2	64.948 ± 0.647	94.456 ± 0.860	75.755 ± 0.815
Midjourney	81.028 ± 0.911	71.074 ± 0.926	85.819 ± 1.249
ERNIE-ViLG	90.243 ± 0.768	232.666 ± 2.742	110.253 ± 1.379

additional model, ERNIE-ViLG. Since we were not able to generate 5k face images with ERNIE-ViLG, we sampled with replacement. We also tried to compared these models against face images we extracted from COCO (without filtering bad images), and the FID scores were extremely high. The FID score between real face images from GFW and COCO (without manual filtering) was 41.23 ± 0.53, which indicates that it is not a good dataset to compare against. We computed FID scores for eyes and mouth using 1k images instead of 5k.

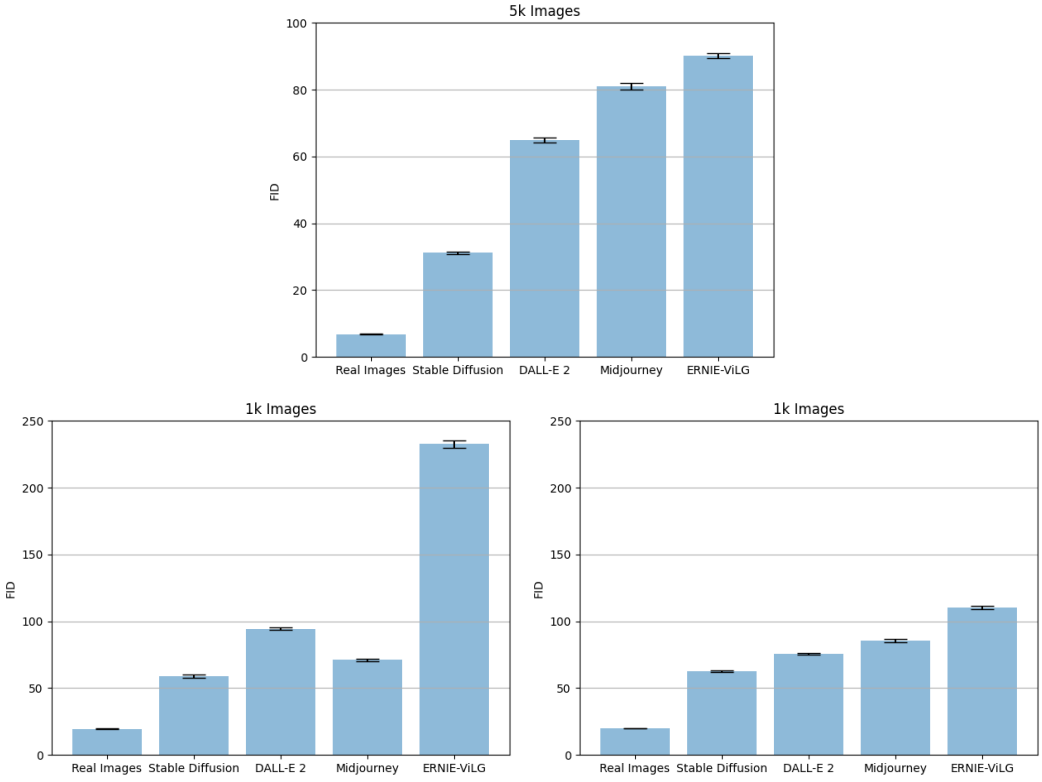

Figure 1: Top: FID scores of models over random sets of 5000 faces. Bottom left: FID scores of models over random sets of 1000 eyes images. Note that we only have 308 eyes images from ERNIE-ViLG. Bottom right: Result with mouth images. Notice that the lower the FID, the better. Results are averaged over 10 runs.

2.4 Results

FID scores on face images from Borji [1] are reproduced and shown in Figure 1, as well as the results on facial details, eyes and mouth. A summary of all results is shown in Table 1.

Figure 1 suggests that these models are generally better at depicting the mouth than the eyes. Even though ERNIE-ViLG has the highest FID scores in all three categories we examined, qualitative inspection shows the model is able to generate high-quality faces images that are almost impossible

to distinguish from real face images. See Figure 2. These are the possible reasons why ERNIE-ViLG performs worse than other models.

- Even though the images were generated in "Realistic Style" and included the description "show front face, like a photo in real life" in the prompts, the model still generates anime and drawing-like face images sometimes. ERNIE-ViLG's demo [5] had seventeen styles available, and some preliminary examination indicates that the model is confused about these styles at times. The inception model behind FID is trained on natural images, so the FID scores would be exceptionally high for images that do not look like a photo from real life.
- Due to limited access to ERNIE-ViLG, we were not able to massively generate images, only generated 1506 images and collected 750 face images. The minimum recommended sample size is 10,000 for calculating FID scores.

3 Future work

Below are some possible aspects that may be worth investigating.

- Try to detect faces with MediaPipe (it was used in Borji [1]). Different face detectors may have some effects on the resulting face image.
- A Large number of images are needed for any kind of quantitative analysis, but many of these models have made it impossible in various ways for various reasons. Crowdsourcing may be one possible solution to it.
- We can also try to evaluate models with other evaluation metrics, such as Kernel Inception Distance (KID) Bińkowski et al. [11] which has a simple unbiased estimator and is believed to work well with a small dataset.

References

[1] Ali Borji. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*, 2022.

[2] Alberto Romero. Dall·e 2, explained: The promise and limitations of a revolutionary ai. URL `https://towardsdatascience.com/dall-e-2-explained-the-promise-and-limitations-of-a-revolutionary-ai-3faf691be220`.

[3] ELIZA STRICKLAND. Dall-e 2's failures are the most interesting thing about it. URL `https://spectrum.ieee.org/openai-dall-e-2`.

[4] Swimmer963. What dall-e 2 can and cannot do. URL `https://www.lesswrong.com/posts/uKp6tBFStnsvrot5t/what-dall-e-2-can-and-cannot-do`.

[5] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022.

[6] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *arXiv preprint arXiv:2210.15257*, 2022.

[7] 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.

[8] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[5] https://huggingface.co/spaces/PaddlePaddle/ERNIE-ViLG

[9] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503, 2016.

[10] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, August 2020. Version 0.2.1.

[11] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

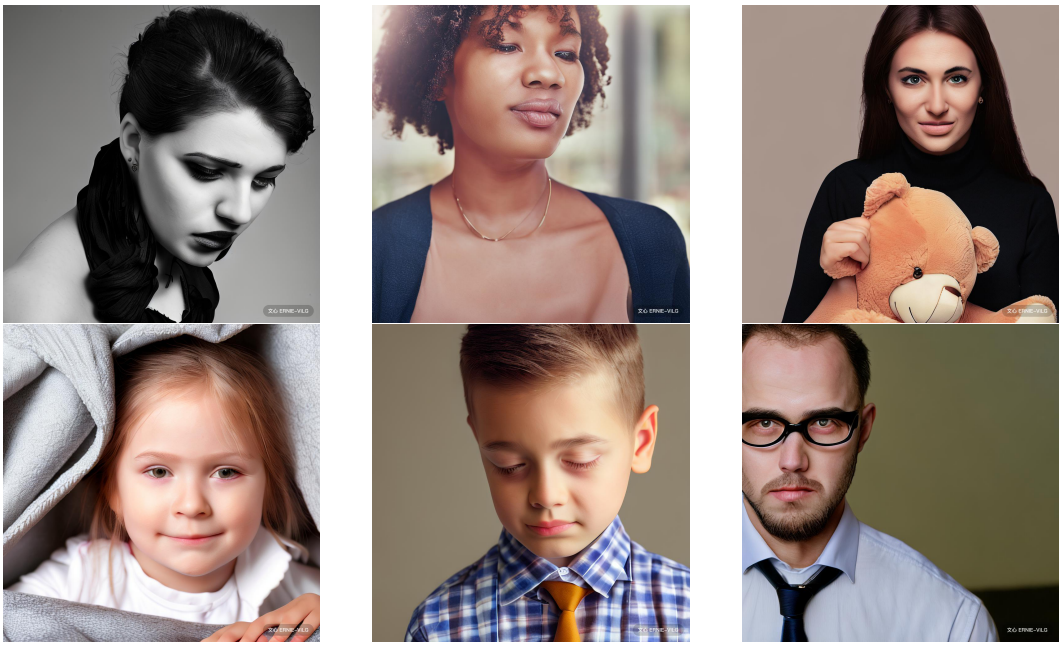

Figure 2: Some of face images generated by ERNIE-ViLG