

Generating Synthetic Satellite Imagery for Rare Objects: An Empirical Comparison of Models and Metrics

Tuong Vy Nguyen^{1,†}, Johannes Hoster^{1,†}, Alexander Glaser^{2,3}, Kristian Hildebrand¹ and Felix Biessmann^{1,2,*}

¹Berlin University of Applied Sciences (BHT), Germany

²Einstein Center Digital Future, Berlin, Germany

³Program on Science and Global Security, Princeton University, New Jersey, USA

Abstract

Generative deep learning architectures can produce realistic, high-resolution fake imagery – with potentially drastic societal implications. Assessing the risks of this technology for the general public requires better understanding of the conditions under which novel generative methods can generate realistic data. A key question in this context is: How easy is it to generate realistic imagery, in particular for niche domains. The iterative process required to achieve specific image content is difficult to automate and control. Especially for rare classes, it remains difficult to assess *fidelity*, meaning whether generative approaches produce realistic imagery and *alignment*, meaning how (well) the generation can be guided by human input. In this work, we present a large-scale empirical evaluation of generative architectures which we fine-tuned to generate synthetic satellite imagery. We focus on nuclear power plants as an example of a rare object category - as there are only around 400 facilities worldwide, this restriction is exemplary for many other scenarios in which training and test data is limited by the restricted number of occurrences of real-world examples. We generate synthetic imagery by conditioning on two kinds of modalities, textual input and image input obtained from a game engine that allows for detailed specification of the building layout. The generated images are assessed by commonly used metrics for automatic evaluation and then compared with human judgement from our conducted user studies to assess their trustworthiness. Our results demonstrate that even for rare objects, generation of authentic synthetic satellite imagery with textual or detailed building layouts is feasible. However, in line with previous work, we find that automated metrics are often not aligned with human perception – in fact, we find strong negative correlations between commonly used image quality metrics and human ratings. We believe that our findings enable researchers to better assess the strengths and weaknesses of different generative methods, especially for niche domains and rare object classes, and can help guide future improvements of generative methods.

Keywords

generative AI, synthetic data, satellite imagery, human evaluation

1. Introduction

With the advent of novel generative methods for tabular data [1, 2], text [3] and images [4], synthetic data has entered the main stage of machine learning (ML) research. The applications are manifold, ranging from arts over software development to improving ML itself.

For generative Artificial Intelligence (genAI), methods designed for text data, the risks and societal impact have been studied, for instance, in the context of large election campaigns [5]. For the image domain however, research on the technical underpinnings of the risks for the general public inherent to genAI technology have been underrepresented in the literature.

Within the ML community, the primary use case for synthetic data is arguably the generation of new training data for the development of larger and more powerful



Figure 1: Samples of synthetic satellite imagery of facilities. Generated using fine-tuned generative models with only text input (top row) and the same fine-tuned models with additional image input (bottom row).

generative ML models. This application scenario has attracted attention, especially in the context of tech companies' demand for more data. These companies could soon run out of data for training language models [6]. Synthetic data appears to be the solution to these problems of data-hungry large ML models, not only for these applications. Moreover, in other application domains, such as health care, synthetic data has attracted attention for different reasons: patient records are sensitive data which must not be shared publicly, hence synthetic

2nd Workshop on 'Public Interest AI' co-located with the 47th German Conference on AI (KI 2024), September 23, 2024, Julius-Maximilians Universität Würzburg, Germany.

*Corresponding author.

[†]These authors contributed equally.

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)



patient record data could solve the problems around training ML models for healthcare without risking patients' privacy [7].

While the opportunities synthetic data offers are certainly convincing, there are two key challenges: For one, it remains difficult to control the output of these methods, second, it remains difficult to automatically assess the quality of generated data. In user-facing products, these two challenges are usually met with extensive human work – there are highly-paid prompt engineering positions and crowd working platforms for evaluating the quality of generated data. These measures jeopardize the very idea of generative methods, automated data generation, and highlight the need for a better understanding of control and evaluation of generative methods.

In this work, we investigate two main questions around state-of-the-art image generation methods: 1) *How easily can they be controlled?* and 2) *How trustworthy are established automatic evaluation metrics?* Our main contribution is a large-scale empirical evaluation with more than 30,000 human ratings of generated images. In contrast to previous work, we take the evaluated methods outside of their comfort zone: First we consider a niche domain, satellite imagery and second, we focus on a specific class of objects for which only a few hundred real-world instances exist, nuclear plants.

Objective We aim to leverage and investigate ML techniques in combination with a pre-trained text-to-image model to generate synthetic satellite images for remote sensing purposes that could combat the issue of insufficient labeled imagery. With our approach, appropriate data could be easily generated in a controllable setting while maintaining low costs. To examine the generalizing capabilities of the pre-trained model and training on limited data, we generate imagery of a unique and rare class. Given the context, we opt for nuclear power plants as our target object and show how synthetic satellite imagery of a specific class can be created under set conditions. We use text prompts and additional image input as guidance for the generation process, and combine this workflow with fine-tuned models.

Moreover, we conduct an empirical study regarding the comparison of different approaches for conditional image synthesis, as well as the automated evaluation metrics used to assess the generated data. Current established evaluation metrics for the assessment of perceptual image quality are known to not necessarily align with human perception and are, thus, not always a reliable measure of image quality [8]. In this work, we acquire quantitative evaluation results through larger-scale human-in-the-loop experiments, which we then compare to established metrics in the field. To the best of our knowledge, such an in-depth empirical comparison between models, metrics and human ratings has not been done for rare object

classes, such as nuclear facilities, in this manner before.

2. Related Work

In the following section we provide an overview of the current state of research in the relevant fields: Generative modeling, particularly text-to-image synthesis, deep learning (DL) applications in remote sensing, and evaluation methods in the context of genAI.

2.1. Deep Generative Models for Image Synthesis

With advances in ML and DL, the process of image generation can be automated and provides a way to generate data at relatively low cost. There are several popular model architectures that aim to generate data and for a long time, the state of the art in generative modeling have been Generative Adversarial Networks (GAN) [9], which have been explored thoroughly and applied in various domains and used as default for text-to-image translation over the years [10]. More recently, novel approaches to image generation based on Diffusion Models (DM) [11] overcame some of the challenges associated with GAN training, with e.g. Dhariwal and Nichol [12] showing that DMs could outperform GANs in image synthesis. Currently, many generative model architectures, especially text-to-image ones, are based on DMs.

Text-to-image generation is a specific type of generative modeling which combines technologies of two different fields: Computer Vision and Natural Language Processing. Image generation conditioned on text has the advantage that it is very intuitive and easily comprehensible. There exist a variety of such recently developed vision-language models [13, 14, 15, 4, 16], that have already shown how powerful such large text-to-image architectures can be. Especially their zero-shot ability - synthesizing images of concepts not seen during training - makes these models very compelling.

Furthermore, there are attempts to directly manipulate features in the latent space to create desired images. For example, by modelling the independent latent characteristics of an object through disentangled representations so that these features can be edited, e.g. changing pose and appearance respectively [17]. Or by identifying latent directions through PCA, which enables to control GAN model-based image features like viewpoint, aging, lighting, and time of day [18]. Park et al. identify a local latent subspace within the latent space of a diffusion model, which enables image editing capabilities through movement along the basis vector at specific timesteps [19].

2.2. Deep Learning in Remote Sensing

There are a variety of use cases for the utilization of DL approaches in the context of satellite imagery and remote sensing, which, since 2014, several works have already dedicated themselves to [20]. Popular use-cases are scene classification, object detection and segmentation [20]. With the emerging of generative models, GANs have found their way into the domain as well, mostly dealing with image-to-image translation tasks, e.g. translating city styles or creating cartographic representations from satellite images [21, 3]. Tasks such as cloud removal and super-resolution are frequent use-cases as well.

However, there have been few works addressing the synthesis of novel imagery: [22] have implemented and evaluated GANs for the synthesis of aerial imagery, but in an unconditional manner. Others have used special software tools instead of GANs with focus on certain objects and tasks like airplanes [23] and synthetic overhead imagery suitable for building segmentation [24]. While most research aims to generate natural images, some works have already brought the text-to-image approach into the remote sensing domain, enabling the generation of different remote sensing images based on text descriptions [25, 26, 27, 28]. However, these were somewhat restricted by the limited amount of suitable image-text datasets and produced results leaving room for improvement. This work aims to further investigate the conditional generation of satellite imagery of a specific and rare class based on text descriptions by using a pre-trained text-to-image model and limited training data for remote sensing, which, to the best of our knowledge, has only been done in [29] so far. Moreover, the additional use of image input next to text prompts for further conditioning during the image synthesis process for satellite imagery, as done in [30], is underrepresented as well.

2.3. Evaluation of Generative AI

For genAI, there are several ways to evaluate models and their outputs [31]. The most common methods are described in the following:

Automatic metrics. The qualitative evaluation of generative models can be very subjective, and requires adequate quantitative metrics for the systematic assessment of generative models and their synthesized data. However, the evaluation of generative models, or their synthetic data, remains a challenging task: There are no standardized benchmarks or protocols set in place [32, 33], and especially in the domain of satellite imagery, finding suitable datasets for training as well as evaluation is rather difficult, particularly for the specific use case at hand. A lot of evaluation metrics, such as the commonly

used Fréchet Inception Distance (FID) [34], require sufficient real data for comparative analysis, which - like in our case - can be difficult to acquire, due to the fact alone that there are only a few hundred existing nuclear facilities in the world. The Inception Score (IS) [35] only assesses the synthetic images, which, albeit lacking a comparison to real data, is more practicable in our case. Since most metrics rely on the feature space of a pre-trained classification model, this limits the metrics to what the model knows, and renders them biased towards the used Inception model and, thus, the ImageNet dataset which the model was trained on [36]. Another drawback is the need for a large sample size (typically around 50k) to make the metrics robust and reliable, which is not always feasible. Furthermore, the FID is known to not necessarily align with human visual perception [8], especially in the remote sensing and earth observation domain [22]. The automatic, reliable evaluation of generative models and its outputs remains an ongoing research field [31].

Downstream tasks. Another way of evaluating synthetic data is its use in a downstream application task: The generated images can be used in e.g. a classification task to observe how well they are classified by a pre-trained classifier [37]. A second method is to train a classification model on the synthetic data, or parts of it, and then apply the trained model on real unseen data and evaluate based on the predictive performance [31]. However, for this method there has to be suitable real data to test with, which, like in our case, is not necessarily given.

Human evaluation. Another method to assess genAI is human evaluation, with many works resorting to user studies to judge their synthetic data [38, 16, 13, 14, 35]. So-called human-in-the-loop experiments can be a valuable alternative when the aforementioned methods are not applicable or unreliable, and their results are easily comprehensible. Human ratings can give a more accurate assessment for specific tasks when automated metrics fail to reliably capture the image quality. Although there have been works proposing more standardized guidelines for evaluation [8, 33, 39], there are no established protocols set in place for human experiments in genAI, which makes a comparison between published works quite difficult. Moreover, conducting human experiments requires additional work and resources, which are not always at disposal. We aim to conduct human evaluation for a use case, where offline metrics are likely unfitting, and follow recommendations regarding user study settings, to make our findings transparent.

3. Method

By utilizing a pre-trained text-to-image model such as Stable Diffusion and fine-tuning techniques, we are able to leverage its prior knowledge while simultaneously adapting the model to our domain. We have fine-tuned this model on imagery of our target object, nuclear facilities, using DreamBooth [40] and Textual Inversion [41] as fine-tuning methods. For more details, we refer to the respective sources. To test the generalizability and the model’s zero-shot capabilities, we use the unmodified pre-trained model as a baseline to examine how well the prior knowledge can be leveraged to generate satellite imagery of our target object. We then apply the mentioned fine-tuning approaches to further train the model on the datasets described below. For further control during the generation process, we use additional conditioning input with the T2I-Adapter model [42, 30]. Moreover, we combine the two approaches: Instead of using the original Stable Diffusion model as base for the T2I-Adapter, we exchange it with our fine-tuned versions (see Figure 2). The intuition is to leverage the newly learned concepts and use them in synergy with the additional image input. The model might then be more familiar with the given layout and able to generate data that better represents our desired image content.

We have three approaches that rely only on text prompts as input and then all three methods used in combination with the T2I-Adapter, leaving us, in total, with six models to evaluate:

- (1) the original unmodified Stable Diffusion model, *SDiff T2I*,
- (2) the DreamBooth fine-tuned model, *DB T2I*,
- (3) the fine-tuned one using Textual Inversion, *TI T2I*,
- (4) the base model (1) with the T2I-Adapter, *SDiff T2I+Adapter*,
- (5) the model from (2) with the T2I-Adapter, *DB T2I+Adapter*,
- (6) the model from (3) with the T2I-Adapter, *TI T2I+Adapter*.

The implementation in this work relies heavily on Hugging Face’s Diffusers Library [43]. For all approaches, we use the publicly available Stable Diffusion v1.5 as base, a pre-trained vision-language model build on Latent Diffusion Models [4]. This version is compatible with the pre-trained T2I-Adapter components and lays a consistent foundation across methods for later comparison.

3.1. Data

To acquire data to train with, Google Earth Engine and web scraping tools are applied to obtain satellite and

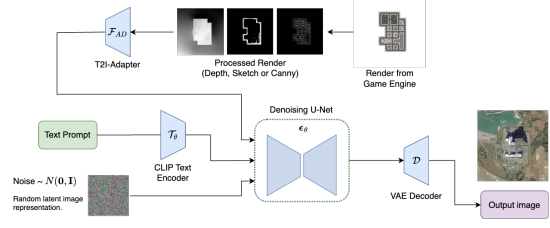


Figure 2: Workflow of our image generation process. We use the renders from the game engine and process them into the respective input modalities (depth/sketch/canny). If used, they’re put into the T2I-Adapter and used as structural guidance in the denoising component, together with the text embeddings obtained from the CLIP text encoder. A text prompt could be “an aerial view of a [*] nuclear power plant”. All components are pre-trained, in case of fine-tuning, the U-Net (with DreamBooth) or text encoder (with Textual Inversion) are modified.

aerial imagery of nuclear facilities. After removing images where sites were blurred or of low quality, the resulting dataset contains 202 satellite images of 185 unique nuclear power plants around the world. To exploit the model’s prior, we apply conditionings - which are not present in the mentioned training data - to those newly learned concepts by adding keywords to the text prompts for variations regarding the location, seasonality and the time of day, for example, generating images of a nuclear facility in the desert or in the winter. Using different settings, synthetic images are generated for each approach. For the additional image input, we use the T2I-Adapter [42] as described in [30]. We generate three different layouts of fictional power plants using the game engine Unity, varying the angle and rotation from which the site is looked at. This creates different viewpoints from the same facility. The renders are then turned into canny edge, depth maps and sketches for further structural guidance during the generation process (see Figure 2). Images are then generated using layout conditioning in addition to the text prompts.

We generate a pool of images for each model, using different variations in the given text prompts. For the additional usage of the T2I-Adapter, we use different input modalities (canny, depth map, sketch) and vary the viewpoints for each of the three layouts. This way, we generate a variety of synthetic imagery, but based on the same three layouts originally rendered from the game engine. For the human experiments, we randomly select 500 images for each approach. Figure 1 shows samples of synthetic satellite imagery of nuclear facilities which have been generated using the methods mentioned previously in this section. For these, either a single text prompt (e.g. “an aerial view of a [*] nuclear power plant, forest, green”) or a text prompt with an additional image have been used as input. For comparison, we also

include the 202 real images in our human evaluation experiments. All images have been scaled to the same pixel size (512x512).

3.2. Experiments

To evaluate our generated data, we conduct a user study where we assess based on three aspects: (a) Fidelity (image quality), how authentic does an image look, (b) text alignment (semantic control), how well does an image match the given text prompt and (c) layout alignment (structural control), how well can the structure of the same subject be retained within several images. For human experiments in genAI, some works opt for a 2-alternative forced choice setup [15, 16, 22]: Two images of two models are put next to each other and the user is tasked with selecting the superior one. However, this user interface design limits the comparison to only two options at a time. Similar to other work [33, 44], we apply a Likert scale where users are tasked to rate a given image or group of images from 1-5, depending on what aspect is being evaluated. This way, images and methods can be rated independently of one another and then later evaluated and ranked.

User study. We use the crowdsourcing platform Toloka [45] and considered two main principles in the experimental design: The task should be simple and the results interpretable [33]. The implemented user interface design for each study is shown in Figure 3: For the (a) image fidelity analysis (see Figure 3a), users are instructed to rate a given image from 1 (unrealistic) to 5 (realistic) based on how authentic it looks to them. The interface for the (b) text alignment studies (see Figure 3b) looks almost the same, with the exception that the text prompt that was used as conditioning input, is also shown. Participants have to rate from 1 (does not match at all) to 5 (matches exactly) how well the shown image matches the text. The third user study examines the (c) layout alignment. As depicted in Figure 3c, the user is shown four images that were generated from the same model, and asked to rate from 1 (completely different facility in each image) to 5 (identical facility) whether the shown images depict the same facility. Participants were selected to be fluent in English and instructed about the motivation of the study; all participants confirmed acceptance of the data usage. We excluded responses from participants that submitted incomplete tasks, tasks in which users pressed only one key repeatedly and tasks in which control tasks (for which ground truth data was available) were answered incorrectly. For tasks (a) and (b) every participant rated 30 images in total. In task (c) each user rated 40 images. Compensation was according to the minimum wage in the most frequent countries of origin on the platform.

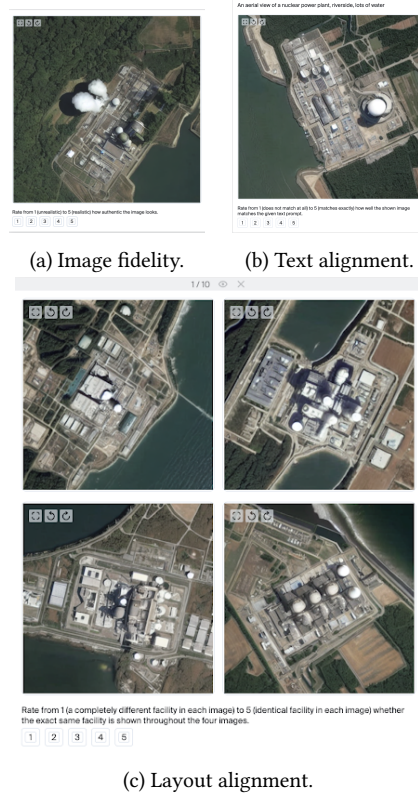


Figure 3: User interfaces of the conducted user studies. Shown are the study designs of the (a) image fidelity, (b) text alignment and (c) layout alignment assessments.

Evaluation setup. Since we lack sufficient real data of our target object, as it is a rare object class, we only consider metrics which do not rely on the feature space of real data. Calculating scores like the FID [34] with the 202 images (see Section 3.1) might deliver unreliable results due to the low sample size and possibly contain bias, as these were already used for fine-tuning. Therefore, we only use the IS [35] as automatic metric in our evaluation, although it also is a flawed metric [46]. For implementation, we use torch-fidelity [47] to calculate the score. The IS is calculated as follows:

$$\text{IS} = \exp(\mathbb{E}_{\mathbf{x} \sim p_\theta} [D_{KL}(p(y|\mathbf{x})||p(y))]). \quad (1)$$

\mathbf{x} is sampled from p_θ , the encoded distribution of our synthetic images. The metric makes use of the KL divergence, calculated between the conditional label distribution $p(y|\mathbf{x})$ (favoring low entropy) and the marginal distribution $p(y)$ from all samples (favoring high entropy). For more details see [35]. To gain a score that might possibly be more accustomed to the remote sensing domain, we further apply an adapted version of the IS, as done in [29]: We exchange the pre-trained Inception model

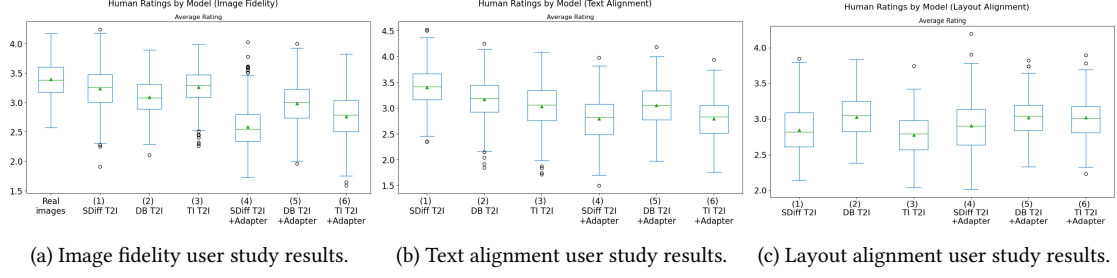


Figure 4: All ratings have been normalized for every user and then scaled back to range 1-5. Human ratings of images generated with different models show robust rankings for (a) image fidelity and (b) text alignment. Results for (c) layout alignment seem less expressive, although slight differences are still visible. Humans rate real images consistently as most realistic, but there are substantial differences between models w.r.t. fidelity and alignment scores.

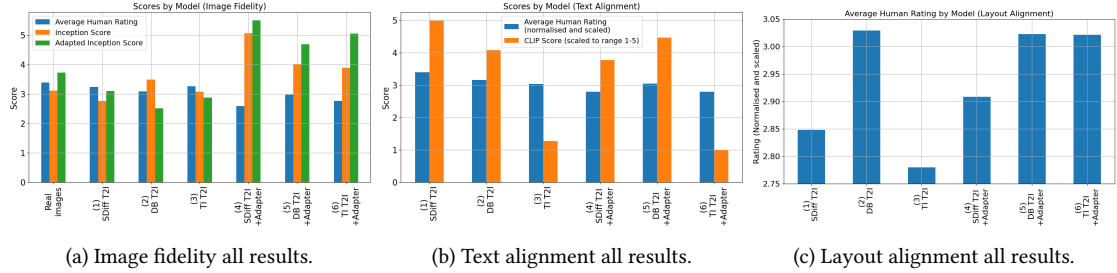


Figure 5: The plots show the quality assessment of images generated with different models. Results are depicted by metric for each model for experiments based on (a) image fidelity, (b) text alignment and (c) layout alignment. A comparison of human ratings and offline metrics demonstrates that there’s a disparity between the two: Image quality of different models is judged differently, images getting a high rating from humans can get a low score from offline metrics.

with a classifier fine-tuned on a land-use classification dataset [48], aerial-view imagery to give information on land cover. The modified IS is denoted as IS_{adapt} .

In regard to the compatibility of image-text pairs, a common metric is the CLIPScore [49], which relies on the CLIP [50] model: It measures the image-text similarity, thus the higher the score, the better. The CLIPScore is defined by the cosine similarity between the image CLIP embeddings c and text CLIP embeddings v (see Equation 2). The score is bound between 0 and 100.

$$\text{CLIPScore}(c, v) = \max(100 * \cos(c, v), 0) \quad (2)$$

For our experiments, CLIPScore is calculated using an open-source package [51], the default CLIP model is ViT-B/32.

For correlation analysis, we investigate the relationship between the automated metrics and the respective human ratings. For this, we calculate the standard Pearson and Kendall Tau correlation coefficients, and the Spearman rank correlation.

Regarding the obtained human ratings, we have to consider that users rate differently than others and might in general be more generous or pessimistic with their judgements. Therefore, we normalize the scores to reduce

the individual user bias and spread: From each available rating x of a user j , we subtract the mean of the ratings of that user, \bar{x}_j , to obtain the normalized value $\tilde{x}_{i,j} = x_{i,j} - \bar{x}_j$. We then scale all ratings back to the original scale of 1-5. For each image, we then calculate the mean rating and aggregate these for each model to obtain an average score for each approach.

4. Results

In the following, we analyze the results of the user study and compare the human ratings with established metrics. The size of our sample after applying the quality control on the collected study results, as described in Section 3, is listed in Table 1. For the (c) layout alignment study, one group of images had to be removed for each model due to technical error, leaving us with a sample size of 496 images for each approach.

Image fidelity. Regarding (a), scores are shown in Table 2. Despite the smaller sample size, we can infer that the real images achieve the best results during the human experiments, followed by the text-only approaches and then the methods combined with the additional image in-

Table 1

Number of ratings, images and participants in user study.
(*From 744 image groups of four.)

	# ratings	# images	# participants
(a) fidelity	16125	3202	447
(b) text alignment	14625	3000	451
(c) layout alignment	3412	2976*	290

put. A visual depiction is shown in Figure 4a. Excluding the real imagery, for the synthetic data the fine-tuned methods yield mostly better results than the corresponding original Stable Diffusion approaches, apart from the (2) *DB T2I* method.

Table 2

Image fidelity results. The real images achieve the best results in human evaluation but receive only average scores with the automated metrics. (*Note the lower sample size compared to the other approaches.)

Model/Data	Sample Size	IS \uparrow	IS _{adapt.} \uparrow	Human Perception \uparrow	
				not normalized	normalized
Real images	202*	3.09 \pm 0.31	3.74 \pm 0.42	3.75\pm0.57	3.39\pm0.29
(1) SDiff T2I	500	2.76 \pm 0.19	3.10 \pm 0.29	3.45 \pm 0.69	3.23 \pm 0.37
(2) DB T2I	500	3.48 \pm 0.29	2.49 \pm 0.24	3.21 \pm 0.61	3.09 \pm 0.30
(3) TI T2I	500	3.04 \pm 0.20	2.85 \pm 0.28	3.50 \pm 0.65	3.26 \pm 0.30
(4) SDiff T2I+Adapter	500	5.05\pm0.40	5.51\pm0.48	2.29 \pm 0.69	2.58 \pm 0.38
(5) DB T2I+Adapter	500	3.98 \pm 0.28	4.72 \pm 0.30	3.02 \pm 0.70	2.98 \pm 0.34
(6) TI T2I+Adapter	500	3.91 \pm 0.24	5.03 \pm 0.55	2.29 \pm 0.75	2.76 \pm 0.39

Using additional input with the T2I-Adapter component gives us more control over the image composition during the generation process, however, the generated images seem to lack image quality: They achieve poorer results than the pure text-based approaches (see Figure 4a). But the fine-tuned approaches (5, 6) yield better results in combination with the layout control in comparison to the original (4) base model. With the text-only approaches, the unmodified model (1) achieves results comparable to the fine-tuned models, but these images often don't show the desired satellite perspective: This aspect is not considered with the Likert scale and was not an influencing factor for the users regarding image fidelity, however this is a limiting factor for the generation of satellite imagery. There was a significant difference between the ratings depending on what was used as input modality (e.g. canny input seems to produce lower human ratings than sketch or depth maps), however this was not further investigated in the scope of this work.

In contrast to the human ratings, automated metrics consistently rank *Adapter*-approaches higher than generative models based solely on text input (see Figure 5a). Furthermore, the real images achieve a relatively average IS and IS_{adapt.} in comparison to the other models.

Text alignment. Evaluation results for (b) text alignment are shown in Table 3. Here, the original model (1) achieves the best image-text alignment scores from human perspective as well as the CLIPScore. Following,

Table 3

For the text alignment results, CLIPScore seems to align with human judgement. (1) achieves the best results in this study.

Model	Sample Size	CLIPScore \uparrow	Human Perception \uparrow	
			not normalized	normalized
(1) SDiff T2I	500	32.74	3.62\pm0.69	3.40\pm0.38
(2) DB T2I	500	31.24	3.28 \pm 0.76	3.17 \pm 0.39
(3) TI T2I	500	26.64	3.00 \pm 0.83	3.03 \pm 0.43
(4) SDiff T2I+Adapter	500	30.74	2.59 \pm 0.81	2.79 \pm 0.43
(5) DB T2I+Adapter	500	31.87	3.05 \pm 0.75	3.05 \pm 0.38
(6) TI T2I+Adapter	500	26.19	2.57 \pm 0.81	2.79 \pm 0.41

the DreamBooth fine-tuned approaches (2, 5) yield the second-best results. Apart from this, the text alignment seems to result in mostly poorer results with the addition of image input, according to our human evaluation, as shown in Figure 4b. However, this ranking does not exactly align with the CLIPScore results. Here, both Textual Inversion fine-tuned approaches (3, 6) significantly perform the poorest (see Figure 5b).

Layout alignment. Since there are no suitable quantitative metrics to evaluate the layout alignment across several images, at least when the viewpoint and conditions are different in each, we only look at the human judgement results for the (c) layout alignment experiments. The results are shown in Table 4: Contrary to expectations, the DreamBooth fine-tuned approach (2) without additional image input achieves the best results in our human evaluation, as also visible in Figure 5c and Figure 4c. However, only by a small margin. One possible reason could be, that the raters might still have been influenced by the image quality or other distortions and details, instead of solely focusing on the layout aspect. In general, the additional conditioning input through the adapter does lead to a better structural alignment according to the scores. Except for (2), all *Adapter*-approaches (4, 5, 6) outperform the ones that are solely based on text input, (1, 3). For the *Adapter*-approaches, fine-tuning seems to help the retaining of the given layout structure during the generation process, as these (5, 6) achieve noticeable better scores than (4).

Table 4

For layout alignment, only human evaluation is available. (2) performs the best, but apart from this, the *Adapter*-approaches are mostly better at structural control.

Model	Sample Size	Human Perception \uparrow	
		not normalized	normalized
(1) SDiff T2I	496	2.656 \pm 0.75	2.848 \pm 0.36
(2) DB T2I	496	2.914\pm0.67	3.029\pm0.30
(3) TI T2I	496	2.416 \pm 0.65	2.780 \pm 0.29
(4) SDiff T2I+Adapter	496	2.688 \pm 0.77	2.909 \pm 0.37
(5) DB T2I+Adapter	496	2.858 \pm 0.61	3.023 \pm 0.29
(6) TI T2I+Adapter	496	2.843 \pm 0.62	3.022 \pm 0.31

Correlation. A correlation analysis has been performed between the quantitative metrics and the respective user study results, the scores are shown in Table 5. For each model, the scores are visually depicted in Figure 6.

For (a) image fidelity, we see that current evaluation metrics, such as the IS and also its adapted version, don’t align with human visual perception. They even correlate negatively, albeit a little less for $IS_{adapt.}$ compared to the original IS, when looking at the correlation coefficients in Table 5. The negative correlation is also visible in Figure 6a. For (b) text alignment, CLIPScore seems to correlate mostly well with the human ratings, as evident by the positive scores in Table 5 and visually in Figure 6b. However, for some models the ranking does not match that of the user studies (see Figure 5b). Furthermore, the (4) *SDiff+Adapter* model achieves a relatively high score as well, although scoring second lowest according to human judgement.

Table 5

Correlation between human judgement and automated metrics. For (a) fidelity, the metrics correlate negatively with the collected ratings. Regarding (b) text alignment, CLIPScore seems to approximate human judgement well.

	Correlation	Pearson	Spearman	Kendall
(a)	Human Rating vs IS	-0.91	-0.82	-0.62
	Human Rating vs $IS_{adapt.}$	-0.79	-0.68	-0.52
(b)	Human Rating vs CLIPScore	0.61	0.89	0.73

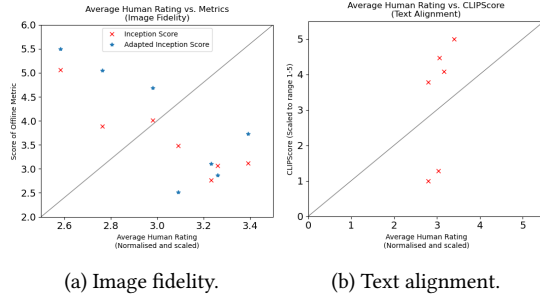


Figure 6: Visual comparison between the automatic metrics and the human ratings (normalized and scaled) from the user studies. Comparison between (a) IS scores and human judgement regarding image fidelity and (b) CLIPScore and human judgement regarding image-text alignment. There is a negative correlation visible between the IS/ $IS_{adapt.}$ and human ratings (a). The CLIPScores and human ratings regarding text alignment correlate positively (b).

5. Discussion

In this work we investigated to what extent modern generative DL methods can be used to generate imagery of

rare objects in niche domains. A special focus in this work was on a comparison of control mechanisms for generative methods. In order to compare different approaches, we leveraged automated quantitative metrics and compared them with human ratings. In extensive empirical evaluations, we demonstrate that novel image generation methods can be used to generate imagery from niche domains and rare objects. Importantly, we find that textual control works well in many cases. But also controlling the image generation with building layouts is feasible, which allows for more fine-grained control.

Inception Score is not aligned with Human Ratings.

A key finding of our empirical results, which is in line with previous studies [52], is that the automated metrics that are used to optimize and evaluate image generation for rare object classes, do not capture image quality as rated by humans. Our results show that exchanging the off-the-shelf classifier models pretrained on ImageNet, which are used as feature extraction backbone in current metrics like the IS, might provide a slightly better metric. However, our results also demonstrate that despite these adaptations to the domain of interest, the IS is not aligned with human visual perception. According to [36], established metrics as used with the default Inception model as backbone might even behave unfair towards diffusion models. Thus, current metrics, as is, are not a reliable measure of performance and image quality when the benchmark is human perception and the goal of these metrics is to actually approximate human judgement. Especially for rare classes and niche domains, as in our case, established metrics are not a trustworthy method of evaluation. Since in such cases, there is not enough real data to calculate additional comparative metrics (such as the FID, KID, Precision and Recall) to get a more robust and broader spectrum of evaluation, the main current established metric for fidelity is the IS, which, as seen in our experiments, is not reliable. Surprisingly, the IS even correlates negatively with human judgement. This appears to suggest a meaningful relationship between these two aspects – but in the opposite direction from what the IS score is intended to measure.

CLIPScore and Human Ratings. As seen in our (b) text alignment results, CLIPScore seems to correlate mostly well with human judgement, there might only be a slight bias towards the (1) original base model, e.g. (4) *SDiff+Adapter* gets a higher CLIPScore while scoring low in the human experiments. This could be due to both the model and the metric relying on CLIP (Stable Diffusion v1.5 uses the pretrained text encoder CLIP ViT-L/14 as conditioning component). The textual inversion fine-tuned approaches, for example, perform the poorest according to CLIPScore. Fine-tuning with this

method trains in the textual embedding space, thus makes changes to the text encoder component, which would reinforce the assumption. Note that while the positive correlation of CLIPScore suggests that this reliably captures human perception, there could be other explanations for this positive result that we cannot rule out based on our user studies alone. For instance, if the generated images always showed the same (or very similar) facilities, this could lead to high scores in our user studies. Heterogeneity of facilities was not enforced in those models that did not use layout inputs. As both the generating models as well as the score use CLIP backbones, such cases could be regarded as a case of overfitting. This problem has been widely acknowledged: the more expressive the models, the more difficult benchmarking becomes as it can be difficult to ensure a clean train-test split [53].

For (c) structural alignment, the approaches using additional image input specifying the precise layout of a facility would, mostly outperform the ones using only text conditioning. This holds true for most of the methods except for (2) *DB T2I*, which scores highest in terms of alignment – but uses only textual input and no layout input (see Table 4). The reasons for this are unclear, one explanation could be that the user study design was not adequate enough to measure the structural control, or the instructions were not clear enough. Since we only have the human ratings to interpret, comparative analysis to other metrics is not possible, thus, a broad and robust evaluation is difficult. The layout alignment - with still considering different rotations and angles from the same target object - is an aspect which has not been thoroughly evaluated in literature yet and could be investigated in further research.

Looking at the evaluation of generative models and their outputs, the question is also raised whether human judgement should be the standard for assessing image quality: Human perception should be used as benchmark when the goal is to approximate this through a metric. However, approximating human perception is not necessarily relevant for all use cases, as also raised in other work [36]. In fact, some of the most interesting application scenarios, such as generating new training data for ML models, do not involve human perception directly.

Societal implications In the context of public interest, human-centric approaches and human-in-the-loop methods are important to understand the risks of ML technology [54, 55]. For instance, previous work in computer vision has demonstrated that it is possible to decouple human perception from machine perception with synthetic imagery [56]. Complementing this work on divergent perception between humans and machines, our findings show a misalignment between human visual perception and automated evaluation, which provides a flawed foundation and benchmark for future develop-

ment of novel ML technologies. Albeit the participants in our studies are no experts in the domain, they are representatives of the general public. Given fabricated or generated imagery, spreading misinformation is even easier when people are not familiar with such niche content. Works addressing information manipulation via ML and crowdsourcing have gained traction within the “AI for Social Good” community [54], possibly due to the rise of fake news, deepfakes and their effortless distribution through mass media. With easily available technology, virtually anyone can produce synthetic imagery that can, as shown in our experiments, fool the average user. Systems being open to validation by, e.g. being open-source, is necessary for transparency [57]. However, an ill-intended user could use such powerful open-source technology for malicious purposes. In our studies, generated images look authentic enough for users to assume they are real, which, depending on the content, could have implications of public interest for citizens.

6. Conclusion

In this work, we have leveraged a pre-trained vision-language model and fine-tuned it to generate synthetic satellite imagery of a rare object class, which has been underrepresented in literature before. Moreover, we conducted large-scale human-in-the-loop experiments to measure human judgement and compared it with established metrics in the field. We found that additional image input mostly gives more control over the image composition, however, it still remains very difficult to control specific details and generate images of the same exact object with the presented conditioning methods. Our results demonstrate that fine-tuning can help generate imagery of specific images and target objects that are on par with data generated from the original base model, in terms of perceived image quality, but that are more suitable for the remote sensing domain and better display the desired satellite perspective. Consistent with previous works, we confirm that established state-of-the-art metrics to evaluate synthetic imagery do not necessarily align with human perception, at least regarding image fidelity. Our findings show that the IS and its adapted version even correlate negatively. CLIPScore seems to work fairly well for measuring image-text alignment, but might be biased towards models based on CLIP. Overall, we find that large-scale user studies are needed to assess synthetic data in regard to human perception, especially for rare classes, where a broad variety of automated evaluation metrics is not available.

For future work, these experiments could be conducted on an even larger scale and for various datasets also in other domains, to investigate whether the findings of this work generalize to other use cases. In line with previous

work, our results provide empirical evidence that current established metrics do not work well for measuring human judgement, especially for rare objects and domains that contain imagery dissimilar to natural images. The quantitative evaluation with automated metrics in genAI requires a more in-depth study and remains an open field of research, not only for the sake of evaluation itself: Better understanding of the perceived image quality will enable researchers to improve generative models in the future.

Acknowledgments

This research is funded by the German Foundation for Peace Research (Deutsche Stiftung Friedensforschung) and is part of the project “Citizen-Based Monitoring for Peace & Security in the Era of Synthetic Media and Deep-fakes” (FP 06/22 | 01/21- FB3-AdD-Pro). Felix Bießmann received funding from the Einstein Center Digital Future, Berlin and the German Research Foundation (DFG) - Project number: 528483508 - FIP 12.

References

- [1] J. Fonseca and F. Bacao, “Tabular and latent space synthetic data generation: a literature review,” *Journal of Big Data*, vol. 10, no. 1, p. 115, Jul. 2023.
- [2] C. Hassan, R. Salomone, and K. Mengersen, “Deep Generative Models, Synthetic Tabular Data, and Differential Privacy: An Overview and Synthesis,” Aug. 2023, arXiv:2307.15424 [cs, stat].
- [3] B. Zhao, S. Zhang, C. Xu, Y. Sun, and C. Deng, “Deep fake geography? when geospatial data encounter artificial intelligence,” *Cartography and Geographic Information Science*, vol. 48, no. 4, pp. 338–352, 2021.
- [4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 10 674–10 685.
- [5] Center for Security and Emerging Technology, B. Buchanan, A. Lohn, M. Musser, and K. Sedova, “Truth, Lies, and Automation: How Language Models Could Change Disinformation,” Center for Security and Emerging Technology, Tech. Rep., May 2021. [Online]. Available: <https://cset.georgetown.edu/publication/truth-lies-and-automation/>
- [6] C. Metz, C. Kang, S. Frenkel, S. A. Thompson, and N. Grant, “How Tech Giants Cut Corners to Harvest Data for A.I.” *The New York Times*, 2024. [Online]. Available: <https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html>
- [7] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, “Synthetic data in machine learning for medicine and healthcare,” *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 493–497, Jun. 2021.
- [8] S. Zhou, M. L. Gordon, R. Krishna, A. Narcomey, L. Fei-Fei, and M. S. Bernstein, “Hype: A benchmark for human eye perceptual evaluation of generative models,” 2019.
- [9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014.
- [10] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, and J. Yang, “Image Synthesis with Adversarial Networks: a Comprehensive Survey and Case Studies,” *Inf. Fusion*, vol. 72, pp. 126–146, 2021.
- [11] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” Dec. 2020, arXiv:2006.11239 [cs, stat].
- [12] P. Dhariwal and A. Q. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” in *Advances in Neural Information Processing Systems 34, NeurIPS 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 8780–8794.
- [13] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 16 784–16 804.
- [14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” 2022.
- [15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-Shot Text-to-Image Generation,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8821–8831.
- [16] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, S. K. S. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” in *NeurIPS*, 2022.
- [17] P. Esser, J. Haux, and B. Ommer, “Unsupervised robust disentangling of latent characteristics for image synthesis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

- [18] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9841–9850.
- [19] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of riemannian geometry," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 24 129–24 142.
- [20] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, Jun. 2019.
- [21] C. Xu and B. Zhao, "Satellite Image Spoofing: Creating Remote Sensing Dataset with Generative Adversarial Networks (Short Paper)," in *10th International Conference on Geographic Information Science (GIScience 2018)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), S. Winter, A. Griffin, and M. Sester, Eds., vol. 114. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018. ISBN 978-3-95977-083-5. ISSN 1868-8969 pp. 67:1–67:6.
- [22] M. Yates, G. Hart, R. Houghton, M. Torres Torres, and M. Pound, "Evaluation of synthetic aerial imagery using unconditional generative adversarial networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 231–251, 2022.
- [23] J. Shermeyer, T. Hossler, A. V. Etten, D. Hogan, R. Lewis, and D. Kim, "RarePlanes: Synthetic Data Takes Flight," 2020.
- [24] F. Kong, B. Huang, K. Bradbury, and J. M. Malof, "The Synthinel-1 dataset: a collection of high resolution synthetic overhead imagery for building segmentation," 2020.
- [25] M. B. Bejiga, F. Melgani, and A. Vascotto, "Retro-Remote Sensing: Generating Images From Ancient Texts," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, vol. 12, no. 3, pp. 950–960, 2019.
- [26] C. Chen, H. Ma, G. Yao, N. Lv, H. Yang, C. Li, and S. Wan, "Remote Sensing Image Augmentation Based on Text Description for Waterside Change Detection," *Remote. Sens.*, vol. 13, no. 10, p. 1894, 2021.
- [27] R. Zhao and Z. Shi, "Text-to-Remote-Sensing-Image Generation With Structured Generative Adversarial Networks," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [28] Y. Xu, W. Yu, P. Ghamisi, M. Kopp, and S. Hochreiter, "Txt2Img-MHN: Remote Sensing Image Generation from Text Using Modern Hopfield Networks," *CoRR*, vol. abs/2208.04441, 2022.
- [29] T. V. Nguyen, A. Glaser, and F. Biessmann, "Generating synthetic satellite imagery with deep-learning text-to-image models – technical challenges and implications for monitoring and verification," in *IN-MM/ESARDA Joint Annual Meeting*, Vienna, Austria, 2023.
- [30] J. Hoster, S. Al-Sayed, F. Biessmann, A. Glaser, K. Hildebrand, I. Moric, and T. V. Nguyen, "Using game engines and machine learning to create synthetic satellite imagery for a tabletop verification exercise," Vienna, Austria, 2023.
- [31] A. Bandi, P. V. S. R. Adapa, and Y. E. V. P. K. Kuchi, "The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges," *Future Internet*, vol. 15, no. 8, 2023.
- [32] K. Man and J. Chahl, "A Review of Synthetic Image Data and Its Use in Computer Vision," *Journal of Imaging*, vol. 8, no. 11, p. 310, Nov. 2022.
- [33] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh, "Toward verifiable and reproducible human evaluation for text-to-image generation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, 2017, pp. 6626–6637.
- [35] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, Barcelona, Spain*, 2016, pp. 2226–2234.
- [36] G. Stein, J. C. Cresswell, R. Hosseinzadeh, Y. Sui, B. L. Ross, V. Villicroze, Z. Liu, A. L. Caterini, J. E. T. Taylor, and G. Loaiza-Ganem, "Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models," 2023.
- [37] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains," Oct. 2022, arXiv:2210.04133 [cs].
- [38] M. Ding, W. Zheng, W. Hong, and J. Tang, "Cogview2: Faster and better text-to-image generation via hierarchical transformers," 2022.
- [39] C. M. Funke, J. Borowski, K. Stosio, W. Brendel, T. S. A. Wallis, and M. Bethge, "Five points to check

- when comparing visual perception in humans and machines,” *Journal of Vision*, vol. 21, no. 3, p. 16, Mar. 2021.
- [40] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,” *CoRR*, vol. abs/2208.12242, 2022.
- [41] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, “An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion,” *CoRR*, vol. abs/2208.01618, 2022.
- [42] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, “T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models,” 2023.
- [43] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf, “Diffusers: State-of-the-art diffusion models,” <https://github.com/huggingface/diffusers>, 2022.
- [44] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, “Imagereward: Learning and evaluating human preferences for text-to-image generation,” 2023.
- [45] P. N., S. I., U. D. CrowdSpeech, and V. DIY, “Benchmark dataset for crowdsourced audio transcription,” 2021, <https://toloka.ai/>.
- [46] S. Barratt and R. Sharma, “A Note on the Inception Score,” Jun. 2018, arXiv:1801.01973 [cs, stat].
- [47] A. Obukhov, M. Seitzer, P.-W. Wu, S. Zhydenko, J. Kyl, and E. Y.-J. Lin, “High-fidelity performance metrics for generative models in pytorch,” 2020, version: 0.3.0, DOI: 10.5281/zenodo.4957738. [Online]. Available: <https://github.com/toshas/torch-fidelity>
- [48] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” 2010.
- [49] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, “Clipscore: A reference-free evaluation metric for image captioning,” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.
- [50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” *CoRR*, vol. abs/2103.00020, 2021.
- [51] S. Zhengwentai, “clip-score: CLIP Score for PyTorch,” <https://github.com/taited/clip-score>, March 2023, version 0.1.1.
- [52] Y. A. Kolchinski, S. Zhou, S. Zhao, M. L. Gordon, and S. Ermon, “Approximating human judgment of generated image quality,” *CoRR*, vol. abs/1912.12121, 2019.
- [53] A. e. a. Srivastava, “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models,” Jun. 2023, arXiv:2206.04615 [cs, stat].
- [54] Z. R. Shi, C. Wang, and F. Fang, “Artificial intelligence for social good: A survey,” *CoRR*, vol. abs/2001.01818, 2020.
- [55] L. Floridi, J. Cows, T. C. King, and M. Taddeo, “How to design ai for social good: Seven essential factors,” *Science and Engineering Ethics*, vol. 26, no. 3, p. 1771–1796, Apr. 2020. doi: 10.1007/s11948-020-00213-5.
- [56] A. Athalye, L. Engstrom, A. Ilyas, and K. Kevin, “Synthesizing robust adversarial examples,” in *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 1, Jul. 2018. ISBN 978-1-5108-6796-3 pp. 449–468. [Online]. Available: <http://arxiv.org/abs/1707.07397>
- [57] T. Züger and H. Asghari, “Ai for the public. how public interest theory shifts the discourse on ai,” *AI & SOCIETY*, vol. 38, no. 2, p. 815–828, Jun. 2022. doi: 10.1007/s00146-022-01480-5.