Evaluating the Correctness of Text-to-Image Generations

Haoyu Guan, Qixuan Xiao, Honghu Luo, Yanyu Ren

Abstract

As text-to-image synthesis methods being widely used in real-world applications, the need for evaluation metrics is becoming increasingly pressing. In recent years, Inception Score (IS)(Salimans et al., 2016), Fréchet Inception Distance (FID)(Heusel et al., 2017), R-precision, and Semantic Object Accuracy (SOA)(Hinz et al., 2019) have been the popular evaluation metrics used by the SOTA text-to-image synthesis models. Nevertheless, these evaluation metrics only focus on image quality, diversity, and consistency which is not comprehensive. In this project, we propose 2 different methods to evaluate the physical consistency of the image. One method combines segmentation and Vision Transformer (ViT)(Dosovitskiy et al., 2020) to predict and classify the image. Another method fine-tunes CLIP based on physical rules set to learn an image encoder that can be used for scoring and classifying images. We demonstrate that both methods can reach a good accuracy on the dataset we built and can give out a reasonable score to an image.

1 Introduction

In recent years, generative models have acquired the capability to generate natural language that is comparable to human language, create limitless synthetic images of high quality, and produce highly diverse human speech and music. These models can be utilized in various applications, such as generating images from text inputs or learning valuable feature representations. Some state-of-the-art models, like GANs and diffusion models, can generate high-quality pictures on most image-generation tasks.

Despite the rapid growth of text-to-image synthesis methods, current evaluation methods are far from perfect. It is necessary to propose a more comprehensive evaluation framework. Traditional evaluation methods such as Inception Score (IS)(Salimans et al., 2016) and Fréchet Inception Distance (FID)(Heusel et al., 2017) are intuitive but have limited performance. R-precision and Semantic Object Accuracy (SOA)(Hinz et al., 2019) are better as they take the meaning of the text into consideration. Counting Alignment (CA)(Dinh et al., 2021) can evaluate whether the number of objects is correct, but it cannot detect some features that violate physical laws.

Those methods only focus on image quality, diversity, and consistency which are not comprehensive. To make evaluation metrics more comprehensive, We suggest two distinct approaches for assessing the physical coherence of the image. The first approach involves using segmentation and Vision Transformer to predict and categorize the image. It segments the human in an image into different parts first and then uses ViT to classify these preprocessed images and give scores. The second method fine-tunes CLIP based on physical principles to learn an image encoder that can be utilized for scoring and classifying images.

We reproduced several experiments using four different evaluation metrics: Inception Score, Fréchet Inception Distance, Structure of Appearance, and Pixel Accuracy, on four different text-to-image synthesis models: AttnGAN, AttnGAN++, CPGAN, and real images. Using these results as our baseline, we trained and fine-tuned our models. We demonstrated that both techniques can achieve high accuracy on their constructed dataset and can provide a reliable score for an image.

2 Related Work

CLIP To analyze inputs and outputs in a text-to-image model, here introduces CLIP(Radford et al., 2021)(Contrastive Language-Image Pre-training). State-of-the-art computer vision systems are trained to predict a set of object categories. But this type of system restricted generality and usability since demands on supervision is expanded. Natural Language Processing is used to analyze the meaning of text with probability models. By mapping raw text to image, CLIP can predict image captions as visual concepts. It uses an efficient and scalable way to learn SOTA image representations on a data set of 400 million (image, text) pairs. Further, CLIP has been tested performances on various downstream vision tasks, including zero-shot, segmentation, caption, video, etc. As one of its downstream tasks, comparing caption of an image and input text can be used to evaluate their matches.

Capture Sub-parts of Objects Text-to-image generation methods can produce high-quality and high-resolution images, but they restricted on creating contents that human wouldn't accepted. Judge, Localize, and Edit(Park et al., 2022) aims to automatically judge the immorality of synthesized images and manipulate images into a moral alternative. They trained an auxiliary text-based immorality classifier with 13,000 textual examples and corresponding binary labels, and utilized CLIP to convert texts and images into joint embedding, then the recognizer will classified input texts in a zero-shot manner. Next, they extended the textual immorality classifier to visual attribute identification. Employing a random input approach can measure the importance of an image region by setting it masked or observed based on model's decision to classify immorality. By utilizing the idea of textual and visual concepts identification, human information or body parts can be retrieved.

ViT ViT (Vision Transformer)(Dosovitskiy et al., 2020) is a type of neural network architecture that has been shown to perform well on computer vision tasks such as image classification and object detection. It is based on the Transformer architecture originally developed for natural language processing and replaces the traditional convolutional layers with self-attention mechanisms that allow the network to attend to different parts of the input image. This makes it particularly effective for processing large images and handling long-range dependencies. Vit has achieved state-of-the-art performance on several benchmark datasets and is considered a promising direction for future research in computer vision.

Evaluation Metrics Although the great achievements of the state-of-the-art methods for text-to-image synthesis such as GANs, Stable Diffusion, the present evaluation methods are not as desired. The current evaluation pipelines mainly focus on two aspects: the image quality and the conformity between the image and its caption. Some commonly used evaluation metrics for the image quality are Inception Score (IS)(Salimans et al., 2016) and Frechet Inception Distance (FID)(Heusel et al., 2017). IS metric uses the pretrained Inception-v3 model to calculate the Kullback-Leibler divergence (KL-divergence) between conditional distribution and cmarginal distribution of the generated images. FID calculates the Frechet distance between the actual images and the generated images using the feature from the pretrained Inception-v3 model .. In addition to these, many evaluation metrics have been proposed for text-image consistency. R-precision (RP)(Xu et al., 2017) used synthesized image query again the input caption and calculated matching score using cosine similarity between image encoding vector and text encoding vector. Semantic Object Accuracy (SOA)(Hinz et al., 2019) using the pre-trained object detector to evaluate whether objects mentioned in the caption are contained in the image, which ranks the models in a similar way to humans. Furthermore, there are some pipelines that combine different evaluation metrics together to achieve a better performance such as TISE (Text-to-Image Synthesis Evaluation)(Dinh et al., 2021).

CDCL-human-part-segmentation Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation (Lin et al., 2020), is a human body part segmentation method proposed by Kevin Lin and his team. This approach takes advantage of the rich and realistic variations of the real data and the easily obtainable labels of the synthetic data to learn multi-person part segmentation on real images without any human-annotated labels. Without any human labeling, this method performs comparably to several state-of-the-art approaches which require human labeling on Pascal-Person-Parts and COCO-DensePose datasets. Their pre-trained model predicts 6 body parts in the images and achieves 72.82% mIOU on the PASCAL-Person-Part dataset. The segmentation of this model is based on the human skeleton (pose) representation and is less disturbed by other factors such as clothing. The segmentation of the target image will help us to train the classification model later.

3 Preliminaries

Traditional metrics such as IS and FID are used to evaluate image quality. The formulas of IS and FID are defined as follows:

$$IS = \exp(\mathbb{E}_x D_{KL}(p(y|x) \,||\, p(y))$$

$$FID = ||\mu_r - \mu_g||^2 + \mathrm{trace}(\Sigma_\mathrm{r} + \Sigma_\mathrm{g} - 2(\Sigma_\mathrm{r}\Sigma_\mathrm{g})^{\frac{1}{2}})$$

where x is the generated image and y is the class label, $X_r \sim \mathcal{N}(\mu_r, \Sigma_r)$ and $X_g \sim \mathcal{N}(\mu_g, \Sigma_g)$ are the features of real images and generated images extracted by a pre-trained Inception-v3 model. For IS, smaller $P(y|x)$ means the object in the image is more distinct, and larger $p(y)$ means the images are more diverse. For FID, a lower the distance between real images and generated images means better image quality and diversity. Other metrics focus on the consistency between text and image. Semantic Object Accuracy (SOA) is proposed to determine whether the objects in the text can be matched in the image. There are two types of SOA metrics which are SOA-I (average recall between images) and SOA-C (average recall between classes), their formulas are as follows:

$$SOA - C = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|I_c|} \sum_{i_c \in I_c} YOLOv3(i_c)$$

$$SOA - I = \frac{1}{\sum_{c \in C} |I_c|} \sum_{c \in C} \sum_{i_c \in I_c} YOLOv3(i_c)$$

where C is the object class set, I_c is a set of images belonging to object class c and $YOLOv3(i_c) \in \{0, 1\}$ will return 1 if YOLOv3 detected an object of class c. Despite SOA can match objects between texts and images, it fails to consider the relation between objects. Positional Alignment(PA) is proposed to evaluate the position relation between objects. PA defines a set of positional words as W and constructs a query problem. For each generated image G_i and text T_i, it generates mismatched texts F_i by replacing the position word w. In this way, a set $D_w = \{(G_{wi}, T_{wi}, F_{wi})\}_{i=1}^{N_w}$ is created, where N_w is the number of texts having position word w. PA is calculated by the query success rate of triplets in D_w, the formula is as follows:

$$PA = \frac{1}{|W|} \sum_{w \in W} \frac{k_w}{N_w}$$

where k_w is the number of success cases, and $|W|$ is the total number of words. Despite the aforementioned metrics have covered wide aspects, there are more details needed to be considered when we evaluate the text-to-image synthesis. Inspired by the existing metrics, we propose a more comprehensive metric that can evaluate whether the generated images obey the defined physical rules or commonsense which are not mentioned in the original text.

4 Physical Consistency Evaluation and Classification

Inspired by popular state-of-the-art methods for text-to-image synthesis, our approach classifies the physical inconsistency of output images. One approach, we use CLIP to embed image captions and pixels into a common space and assign body words with high prediction to each cluster of pixels. Then, we take image feature encoding to a classification network to produce evaluation metrics. In another approach, we trained a classifier to determine whether the generated single-person images are consistent with physical common sense using CDCL+Pascal Human body part Segmentation+Vision Transformer(ViT).

4.1 Segment+ViT

The core idea of this approach is to use a body part segmentation model to automatically annotate and highlight each part of the human body in the generated images, and later use the ViT model to learn the relative relationships among them.

Because there is a limit to the amount of data that we can label manually and the self-attention layer of ViT lacks locality inductive bias, we need to augment our dataset. We used shift(cv2.warpAffine), RandomRotation(10,90), and flip to augment our data manually. By using data augmentation, we want our ViT model to focus attention on the relative relationship of body parts, rather than memorizing the absolute position of each part on the image. We then use the pre-trained CDCL-human part segmentation model to automatically segment and annotate the body into 7 parts from the generated images. The segmented images are then resized and later used for training the ViT classification model.

Since ViT models require a huge amount of data to achieve good performance. It's not feasible to train a ViT model from scratch. So we used a pre-trained ViT model vit-base-patch16-224-in21k, which was trained on ImageNet-21k(14 million images, 21,843 classes) at resolution 224x224. The pre-trained model learns an inner representation of images that can then be used to extract features useful for downstream tasks. After that, we put in our data (original + augmented) to fine-tune our ViT model. In this way, we obtain a classifier that can determine with acceptable accuracy whether the generated single-person picture conforms to physical common sense.

4.2 Fine-tuned CLIP

As shown in Figure 1, Our model takes advantage of the basic CLIP model. We used three steps to analyze the physical rules of images and evaluate the physical consistency score: 1. generate texts that describe the images, 2. fine-tune the CLIP base model, and 3. classify the image embedding to evaluate scores.

4.2.1 Free-Form Caption Generation

First, we tested the accuracy of the CLIP model with prompts of different structures, content, sensitivity, and inclusiveness. A good finding shows CLIP is not sensitive to the choice of numbers, some words will hint at the entity of images, however, they depend on the quality of data from the pre-trained model. According to each image, we manually annotate them by the following features: how many people are in the picture, the visual impact of character sizes in distances, the direction in which characters are facing, and the correctness of shapes for character head, hands, and legs. Then we use a template to generate free-form captions for the input images, and in addition, on the template, we include the word "human" to imply it's a human-related text-image matching job.

4.2.2 Fine-tuning

Our approach to fine-tuning CLIP for Physical Consistency Evaluation is shown in Figure 1. Specifically, text and image representations are both generated by transformers, vision transformer is applied to produce image representation. The trained image encoder is used to produce evaluation metrics.

Language Encoder We adopt the well-designed pre-trained language model from the CLIP base which is published by OpenAI. We analyzed the language model and found out it has logical flows. And training a language model with 400,000,000 text is difficult for our work due to time limitations, thus we decided to fine-tune the CLIP pre-trained language model. We batched free-form captions into a balanced batch sampler, to maximize and reduce the bias. Then we tokenize batches of free-form captions and feed them into the pre-trained language model, for a total of 10 epochs, and all model weights are updated.

Image Encoder We generated images from captions of the MS-COCO dataset with people objects with the Stable Diffusion model. We leverage the dataset to make our image annotation equally contributed. Then input images are downsampled to be fed into a vision transformer. Assuming the input image size is $H \times W$, and the down-sampling factor is ds, we define $\tilde{H} = \frac{H}{ds}$ and $\tilde{W} = \frac{W}{ds}$.

After the text and image inputs are embedded, we correlate them using inner products, creating a tensor $\tilde{H} \times \tilde{W} \times N$ as the inner product of the N-dimensional vector of text embedding and the image embedding. After obtaining the correlation tensor, we check the cosine similarity of text and image pairs for minimizing it.

4.2.3 Physical Consistency Score Evaluation

For the downstream fine-tuning experiments, we treated the fine-grained physical consistency attributes from the image encoder as a binary classification task where each attribute in an image is assumed as an independent feature and images can be assigned multiple features which are shown in Figure 1. Then we used an MLP layer with a dropout of 0.2 to get the binary classification result. The score is calculated from the weights of matched body parts multiply by the result classification probability.

5 Experiment

In the experiment section, we first test the previous evaluation metrics for text-image matching using the baseline model on the MS-COCO dataset. And some early classification experiments based on whether it conforms to common sense were conducted on hand images. Then we evaluate both the segmentation + ViT method and the fine-tuning CLIP method in the generated images set with people objects from the MS-COCO dataset. We demonstrate the segmentation + ViT method and the fine-tuning CLIP method has a remarkable classification accuracy on our generated dataset. What's more, we will show both methods can give out a reasonable score to judge the physical consistency of the image based on the defined physical rules set.

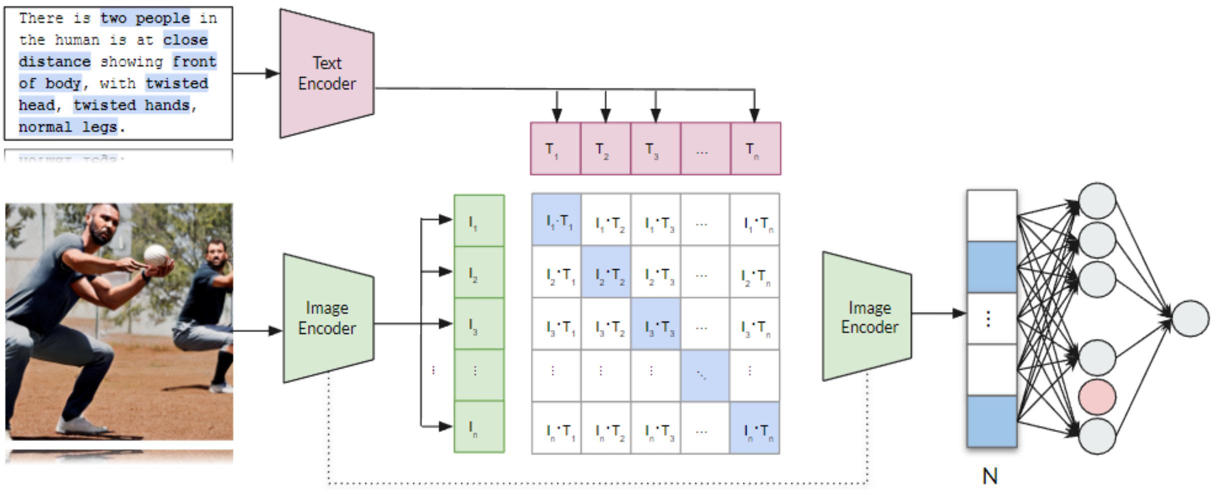

Figure 1: The architecture of Fine-tuned CLIP and Physical Consistency Score Evaluation

5.1 Evaluation Metrics Reproduce

5.1.1 Experimental Setup

Datasets We use the MS-COCO dataset to test the evaluation metrics. This dataset has approximately 120K images, where 80K images are for training and 40K for validation. The MS-COCO dataset also has coordinates of bounding boxes and segmentation masks for 80 categories of objects and pixel maps of 91 categories of background regions like walls, sky, or grass.

Baseline Models We test the current evaluation metrics on some SOTA text-to-image synthesis models. Here we use AttnGAN, AttnGAN++, and CPGAN as the baseline models.

Evaluation Metrics We test the existing evaluation metrics based on the defined dataset and baseline models. We use IS and FID to evaluate the image realism, RP to evaluate the text relevance, SOA to evaluate the object accuracy and PA to evaluate the relation between objects. Here, we use the YOLO-v3 as the object detector to compute SOA.

5.1.2 Results and Discussion

We conduct text-to-image synthesis on the MS-COCO dataset using the baseline models and evaluate them using the evaluation metrics we chose. The result of different metrics on different baseline models is shown in Table 1.

Based on the results, we can draw some insights. Firstly, AttnGAN++ outperforms AttnGAN on all metrics. Secondly, we observe that CPGAN achieves a score close to that of real images, which could be attributed to the use of YOLOv3 in both CPGAN and SOA, leading to potential overfitting.

Table 1: Evaluation Metrics Result

Model	IS	FID	SOA-I/C	PA
AttnGAN	33.76	36.90	49.78/47.13	40.08
AttnGAN++	54.63	**26.58**	69.97/67.83	**47.75**
CPGAN	**59.64**	50.68	**83.83/81.86**	43.28
Real Images	51.25	2.62	91.19/90.02	100

5.2 Early experiments on hand images

5.2.1 Experimental Setup

The first experiment is about determining whether the generated hand images are "true" (in line with physical common sense). The choice of the hand as an experimental target is a first attempt to challenge the current difficulties in the field. At the time we collected the hand dataset, we found that only about 8% of the images generated by Stable Diffusion could be classified as true. It can be said that the current image generation model still cannot generate realistic hand images properly.

Datasets The first part of the dataset consists of 400 generated images of size 512*512 pixels from the stable diffusion official website, with the prompt "single real hand". The second part of the dataset contains a total of 175 real hand images obtained from Adobe Stock. The dataset comprises a total of 575 images, which were later resized to 128*128 pixels to facilitate training and memory for the first experiment. We randomly selected 500 images for the training set and 75 images for the test set.

Evaluation Metrics The determiner is binary, so if an image is considered to be true, it is marked as 1, and if it is false, it is marked as 0. The criteria to label an image as true are that the shape of the hand conforms to common sense, the lines (texture, fingerprint), and the nails of the hand conform to their relative positions and shapes, and the size of each finger is relatively uniform. For the sake of simplicity in the first experiment, we also labeled hands with different colors (stained or lighted) and hands with a small portion of other patterns as correct.

Baseline As we have not been able to find a public model of a "detector that can tell whether a generated picture conforms to physical common sense". We start from scratch, the candidate models are CNN and VIT, and in this initial experiment, we chose the simple CNN model.

5.2.2 Results and Discussion

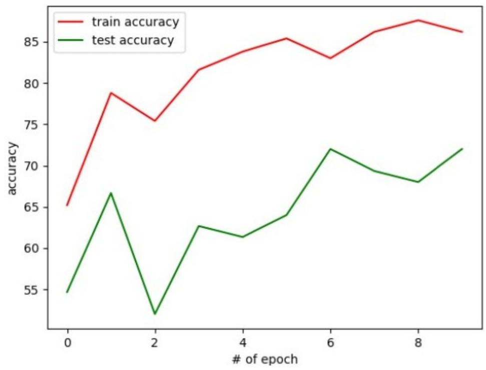

Figure 2: The training and testing acc over epoch

We plotted the relationship between training accuracy and testing accuracy over epochs in Figure 2. We can see that the final training accuracy is not high enough, and there is still a large gap between the test accuracy and the training accuracy. This indicates that our model is not only overfitting but also has extraneous bias interference. This is because our dataset is too small and the generated images generally have darker backgrounds, while a large portion of the true dataset has brighter backgrounds. It is also possible that the simple CNN itself is one of the reasons for the poor training results, and we set the structural complexity of the initial experiments very low. This is not enough for hands with complex features such as shape, texture, relative position, and 3D visual occlusion relations. At a time when it is unable to find hand pictures that are further subdivided and annotated today, the classifier based on Segmentation-learning is difficult to improve on hand images, so we replace our classification objectives from hand to the entire human body.

5.3 Segment+ViT

5.3.1 Experimental Setup

Figure 3: Segmented image sample

Datasets We generated about 2k images (768*768 pixels) from the captions of the MS-COCO dataset with random prompts from a single person, using Stable Diffusion. (For images of plural people, the training results are poor in the current stage of this method.) To facilitate training, we artificially controlled the ratio of good to bad pictures in it to be about 1:1, for EACH pose. For the original images with a high degree of repetition such as "standing", a smaller portion of the dataset should be kept in order to prevent over-fitting. After dividing the dataset into training, validation, and test sets in the ratio of 7:1:2, we used shift(cv2.warpAffine), RandomRotation(10,90), and flip to augment data manually. With data augmentation, we try to make our model learn the relative relationships of various parts of the human body instead of overfitting. We use CDCL+Pascal human body part Segmentation to preprocess the images. We got the segmented image like Figure 3. Finally, we resize them to 224*224 pixels and put them into our ViT model for learning.

Evaluation Metrics Similar to what we did with the hand images. In this experiment, we simply consider whether the person's limbs, head, and torso are present(the obscured part is also considered to be present.) and connected, and whether their number(the three-legged man is certainly not right), relative positions and proportions are consistent with common sense. Details of distortion on the hand and face were ignored in this experiment.

Baseline Model We used the classical ViT model vit-base-patch16-224-in21k, which was pre-trained on ImageNet-21k. Considering the time constraint and the nature of this experiment as a feasibility study, we decided to keep the pre-trained model and fine-tune it using our own data.

Hypothesis Our Hypothesis is: Subdivision and annotation of body parts in generated images will make the training of the model easier. In fact, direct training using the original generated images with a simple ViT model can not give us satisfactory results, the accuracy of the test set cannot be improved, it only over-fits. Compare with the results of our training later using the segmented images, it shows that our hypothesis is relatively correct.

Step	Training Loss	Validation Loss	Accuracy
20	0.606200	0.606544	0.632000
40	0.538100	0.535674	0.784000
60	0.579800	0.488475	0.808000
80	0.405100	0.452231	0.816000
100	0.467800	0.429404	0.832000
120	0.407700	0.396356	0.824000
140	0.325700	0.381295	0.824000
160	0.309600	0.361815	0.832000
180	0.288600	0.354236	0.840000
200	0.221200	0.348060	0.848000
220	0.217700	0.323454	0.864000

Figure 4: The validation accuracy and loss over steps for ViT model

5.3.2 Results and Discussion

The model trained/ tested using all pose prompts can eventually reach 86% training and 85.26% testing accuracy. Even if we completely remove the images of one of the poses from the training set and use all the images of that pose as the test set, we still get a test accuracy of about 76%. This shows that our model has the ability to generalize and reason.

```
***** test metrics *****
  epoch                   =        4.0
  eval_accuracy           =     0.8526
  eval_loss               =     0.3664
  eval_runtime            = 0:00:01.12
  eval_samples_per_second =    168.515
  eval_steps_per_second   =     21.286
```

Figure 5: The test accuracy and loss for ViT model

Most importantly, we validated the idea that continually subdividing, identifying, and learning the relative relationships of parts may be used to make determinations about a wide range of general objects level by level.

```
per_device_train_batch_size=16,
evaluation_strategy="steps",
num_train_epochs=4,
fp16=True,
save_steps=20,
eval_steps=20,
logging_steps=10,
learning_rate=2e-5,
save_total_limit=2,
remove_unused_columns=False,
push_to_hub=False,
report_to='tensorboard',
load_best_model_at_end=True,
```

Figure 6: Hyper parameters for ViT model

5.4 Fine-tuning CLIP

5.4.1 Experimental Setup

Datasets We generated the data from the captions of the MS-COCO dataset with people objects, using the Stable Diffusion model. Our generated dataset has approximately 2500 images, where 2K images are for training and 500 for validation. Since we focused on the human body structure, we defined the physical rules set based on it. Then we labeled each image according to the physical rules set and generated free-form captions of physical rules.

Baseline Methods In experiments, we used ViT-B/32 CLIP as the baseline model to fine-tune. And the visual encoder we learned for the image is ViT-B/32 of CLIP. For the classifier, we use an MLP with a dropout layer.

5.4.2 Results and Discussion

Model Prediction Accuracy The prediction accuracy of our model on the generated dataset is 79.2% as shown in Table 2 which is remarkable. In table 1, we can also see that the classifier can reach a high precision on both 0 and 1 classes. However, it has a poor performance on the recall rate of 0 class which also leads to a poor F1-score. This is possibly due to the data imbalance.

Physical Consistency Score The Physical Consistency Score is calculated from the probability of class 1 which ranges from 0 to 100. As shown in

Table 2: Evaluation Metrics Result

CLASS	Precision	Recall	F1-Score	Support
0	0.82	0.51	0.63	946
1	0.78	0.94	0.86	1554
Accuracy				**0.792**

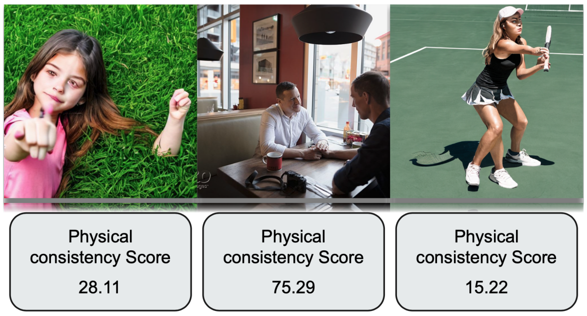

Figure 7: Physical Consistency Score Sample

Picture x, in the first picture and the third picture the girl has a twisted hand and the woman has 3 legs, therefore they both have a low score. The second picture is a normal picture and its score is high. The result shows that our model can rank a reasonable score based on the physical rules set we defined.

6 Conclusion and Future Work

6.1 Conclusion

In this paper, we proposed novel approaches to solve the generalized Physical Consistency Evaluation problem of AI-generated human images. In experiments, we demonstrate that both of our approaches can have a good performance in classification accuracy and give out a reasonable score to judge the Physical Consistency of an image.

6.2 Future Work

Due to the limited time, we are not able to generate and label a large dataset, but a large and balanced dataset would definitely improve our results. Collecting more images that generated different poses and prompts would increase the accuracy.

For the Fine-tuning CLIP method, we focus on the human body structure when defining the physical rules set, future works might further explore whether it can be extended to more generalized physical rules such as the relationship between different objects. Besides, object detection can be utilized to extract foreground objects, which might lead to a more stable result in theory.

For the Segment-ViT method, the idea of segmenting parts and learning relative positions has been proven to work. This idea of continually subdividing, identifying, and learning the relative relationships of parts can be used to make determinations about a wide range of general objects level by level. A tree classification structure can be built, such as segmenting single people from multiple people images, segmenting hands from single people images, and separating thumbs, index fingers, and even nails and fingertips from hands. Then the classification models between layers are determining whether their relative positions, sizes, and numbers match the physical rules and finally determine the whole picture. This requires the labeling of huge amounts of data and the annotation of detailed parts of individual objects. But ultimately, this model can distinguish most of the objects in the world, and widely distinguish whether the images conform to physical common sense. Because this learning process is consistent with the way people think, it will eventually know how to determine whether the whole object is true by the details and the relations of the parts as we do.

Figure 8: Subdivides object parts further

Ethics Statement

We proposed a novel approach to solve the generalized Physical Consistency Score Evaluation problem from AI-generated images. We use public human-related prompts and AI image generation, such as Stable Diffusion to collect data for our experiments. Our code or method is potentially subject to concerns of discrimination/bias/fairness since the current classification of the human body as "normal" is based on the majority of the population, this may lead to potential discrimination against minority groups such as people with disabilities if someone uses it inappropriately. Since our generated images are based on the stable diffusion model, the potential privacy issues associated with the model itself need to be taken into account. However, our results are currently being used only for academic research for non-profit purposes. We are not responsible for any unauthorized use by others that causes ethical problems.

Acknowledgements

This work has been supported by the University of Southern California CSCI 566 TA group in spring 2023. We also acknowledge the computational resources provided by the USC Advanced Research Computing.

A Appendix

The code of our experiments can be found at `https://github.com/Rorschach11/566Project-Evaluating-the-Correctness-of-Text-to-Image-Generations`.

References

Tan M. Dinh, Rang Nguyen, and Binh-Son Hua. 2021. TISE: A toolbox for text-to-image synthesis evaluation. *CoRR*, abs/2112.01398.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2019. Semantic object accuracy for generative text-to-image synthesis. *CoRR*, abs/1910.13321.

Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. 2020. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*.

Seongbeom Park, Suhong Moon, and Jinkyu Kim. 2022. Judge, localize, and edit: Ensuring visual commonsense morality for text-to-image generation.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *CoRR*, abs/1606.03498.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2017. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485.