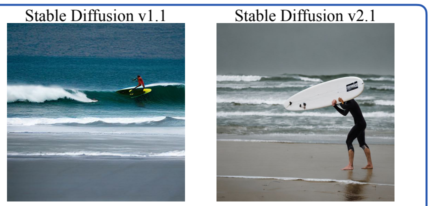
Appendices

A. Qualitative Examples

Here we show three examples on how TIFA is computed. For each text input, we show the questions and answers generated by GPT-3 [3] and filtered by UnifiedQA [25]. We also show the VQA model's answer to each vision question given each generated image. The VQA model used is mPLUG [31]. The first text input comes from COCO [34] and the second and third text inputs come from DrawBench [47].



Text Input: On a gray day a surfer carrying a white board walks on a beach.

GPT-3 generated + verified QAs

Question: is this a surfer?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: who is on the beach?
Choices: surfer, swimmer, driver, skier — Answer: surfer
VQA: Surfer ✓ — Surfer ✓

Question: is the person carrying a board?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: is this a beach?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: what type of place is this?
Choices: beach, park, forest, mountain — Answer: beach
VQA: Beach ✓ — Beach ✓

Question: is the surfer carrying something?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: is the surfer walking on the beach?
Choices: yes, no — Answer: yes
VQA: No ✗ — Yes ✓

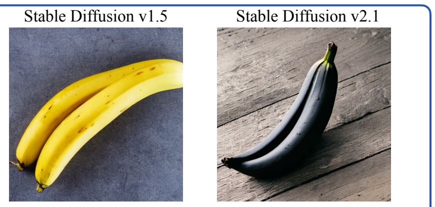
Question: is the day gray?
Choices: yes, no — Answer: yes
VQA: No ✗ — Yes ✓

Question: what color is the sky?
Choices: black, gray, red, blue — Answer: gray
VQA: Blue ✗ — Gray ✓

Question: is the board white?
Choices: yes, no — Answer: yes
VQA: No ✗ — Yes ✓

Question: what color is the board?
Choices: black, white, red, blue — Answer: white
VQA: Black ✗ — White ✓

TIFA — Accuracy on 12 questions
54.6 — 100.0



Text Input: A black colored banana.

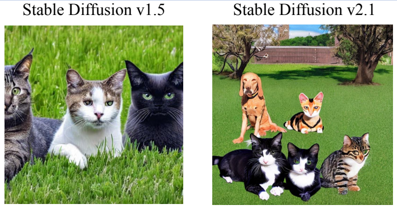
GPT-3 generated + verified QAs

Question: is this a banana?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: is the banana black?
Choices: yes, no — Answer: surfer
VQA: No ✗ — Yes ✓

Question: what color is the banana?
Choices: yellow, red, black, white — Answer: black
VQA: Yellow ✗ — Black ✓

TIFA — Accuracy on 3 questions
33.3 — 100.0



Text Input: Three cats and two dogs sitting on the grass.

GPT-3 generated + verified QAs

Question: are there cats?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: are there dogs?
Choices: yes, no — Answer: yes
VQA: No ✗ — Yes ✓

Question: is there grass?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: what are the animals sitting on?
Choices: grass, leaves, twigs, sand — Answer: grass
VQA: Grass ✓ — Grass ✓

Question: are the animals sitting?
Choices: yes, no — Answer: yes
VQA: Yes ✓ — Yes ✓

Question: how many cats are in the picture?
Choices: 1, 2, 3, 4 — Answer: 3
VQA: 3 ✓ — 4 ✗

Question: are there two dogs?
Choices: yes, no — Answer: yes
VQA: No ✗ — No ✗

Question: how many dogs are in the picture?
Choices: 1, 2, 3, 4 — Answer: 2
VQA: 1 ✗ — 1 ✗

TIFA — Accuracy on 8 questions
62.5 — 62.5

Table 4. Detailed evaluation of each text-to-image model on TIFA v1.0.

	VQA Accuracy by Element Category							
	shape	other	counting	spatial	attribute	activity	food	object
AttnGAN[60]	42.0	47.8	41.9	70.8	53.6	64.3	48.1	56.3
X-LXMERT[5]	34.8	46.8	41.7	70.9	55.2	65.4	52.4	57.0
Stable Diffusion v1.1[46]	66.7	68.7	66.0	69.4	74.6	73.8	79.7	75.1
VQ-Diffusion[16]	63.8	64.2	61.6	73.5	75.4	76.7	80.0	74.2
Stable Diffusion v1.5[46]	65.2	72.1	66.6	72.9	78.0	76.9	81.4	78.4
minDALL-E[27]	**69.6**	**74.6**	69.0	74.7	77.0	79.5	**82.8**	79.9
Stable Diffusion v2.1[46]	66.7	72.1	**73.3**	**76.1**	**78.8**	**82.0**	82.2	**82.4**

					TIFA by text source		**Overall TIFA**
	location	color	animal/human	material	COCO	free-form	
AttnGAN[60]	60.4	56.5	58.6	61.7	67.5	47.4	58.1
X-LXMERT[5]	69.1	54.8	52.8	61.2	68.1	47.7	58.6
Stable Diffusion v1.1[46]	78.4	75.7	78.2	80.4	79.3	72.2	75.7
VQ-Diffusion[16]	77.9	**84.2**	79.0	80.9	79.8	72.6	76.2
Stable Diffusion v1.5[46]	79.9	78.8	80.6	84.7	81.9	74.9	78.4
minDALL-E[27]	82.1	83.7	78.9	86.1	83.5	75.5	79.4
Stable Diffusion v2.1[46]	**82.8**	83.6	**85.2**	**88.5**	**86.3**	**77.7**	**82.0**

B. Detailed Results

Table 4 shows the detailed evaluation results of the text-to-image models we use on the TIFA v1.0 benchmark. We show the VQA accuracy of each question category, TIFA score on each text source, and the overall TIFA score. We can see that Stable Diffusion v2.1 [46] gets the highest overall score and also scores the highest in most categories. Nonetheless, the CLIP [42] and VQGAN [9] based minDALL-E [27] gets the highest accuracy on "shape", "other", "food", and VQ-Diffusion [16] gets the highest accuracy on "color".

C. Annotation Details

C.1. Likert Scale on Text-to-Image Faithfulness

Guidelines The annotation guideline is as follows:

- On a scale of 1-5, score "does the image match the prompt?".
- The ranking of each image given the same text input is important. If you believe the current scoring criteria cannot reflect your ranking preference, pick scores that are consistent with your ranking. Ties are allowed.
- To evaluate the generated image, there are two aspects: image quality and text-image match. Here we only care about text-image match, which is referred to as "faithfulness".
- There are several kinds of elements in the text: object, attribute, relation, and context. Measure the consistency by counting how many elements are missed/misrepresented in the generated image.
- For some elements, e.g. "train conductor's hat", if you can see there is a hat but not a train conductor's hat, consider half of the element is missed/misrepresented in the generated image.
- Objects are the most important elements. If an object is missing, then consider all related attributes, activity, and attributes missing.
- When you cannot tell what the object/attribute/activity/context is, consider the element missing. (e.g., can't tell if an object is a microwave)

Given the above guideline, suppose the text input contains n elements, and x elements are missed or misrepresented. n and x are all counted by the annotators. The reference scoring guideline is as follows:

- 5: The image perfectly matches the prompt.
- 4: $x \leq 2$ and $x \leq n/3$. A few elements are missed/misrepresented.
- 3: $\min\{2, n/3\} < x \leq n/2$ elements are missed/misrepresented.
- 2: $x > n/2$. More than half of the elements are missed/misrepresented.
- 1: None of the major objects are correctly presented in the image.

Details We collect 1600 annotations on 800 generated images from 160 text inputs. Each image is scored by 2 annotators, and we collect the scores from 20 graduate students. We average the scores as the final faithfulness score of the image. The inter-annotator agreement measured by Krippendorf's α is 0.67, indicating "substantial" agreement. The images are generated by the five most recent text-to-image models in our study, including VQ-Diffusion [16], minDALL-E [27], and Stable Diffusion [47] v1.1, v1.5, and v2.1. For each text input, we present the five images together, making it easier for the annotators to give faithfulness scores that reflect their ranking preference. We will release the annotation scores on publication.

C.2. Human VQA

Guidelines Given an image, a question, and a set of choices, choose the correct choice according to the image content. There are two types of questions. One has two options: "(A) yes (B) no". Another type of question has four choices. We also add the fifth option "None of the above". If you believe none of the four choices is correct, choose the fifth one. Some images are of low quality. Just select the choice according to your instinct. For ambiguous cases, for example, the question is "is there a man?", and the image contains a human but it is unclear whether the human is a man, answer "no".

Details We collect annotations of 1029 questions on 126 generated images. The images are from images used in the Likert Scale annotation. Each question is answered by two annotators, and we have the same 20 graduate students as the annotators. The inter-annotator agreement measured by Krippendorf's α is 0.88. A third annotator is involved if the two annotators disagree. And the final answer is given by the majority vote. We will release the annotated VQA answers.

D. Common Q & A

Any possible extension to TIFA? As discussed in §1, one extension of our work will be customized versions of the TIFA benchmark focusing on one aspect of image generation. For example, we can make a TIFA benchmark that only contains questions about "counting"; Or a benchmark consists of text inputs synthesized to test text-to-image models' ability in composing multiple objects. Another possible extension is to use TIFA on other generation tasks, e.g., text-to-3D and text-to-video.

The OpenAI APIs are too expensive. Can we generate questions by local models? Yes. Our approach works on any language model. Please refer to §4.3 on question generation with our fine-tuned LLaMA 2 checkpoint in. Also, we would like to emphasize that all questions in TIFA v1.0 benchmark are pre-generated by GPT-3, and there is no need to re-generate those questions for evaluation.

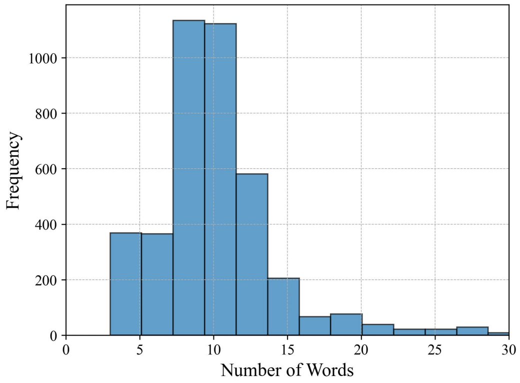


Figure 12. Distribution of the lengths of TIFA v1.0 text inputs.

More details on TIFA v1.0 text inputs? The distribution of the number of words in the text inputs is shown in 12. Most text inputs have around 10 words. We also conduct bias analysis on TIFA v1.0 text inputs. Among the 4K text inputs, regarding gender expression, 400 are perceived as "male" and 239 are perceived as "female". We find that the bias in gender distribution comes from captions sampled from COCO dataset [34].

E. Text-to-Image Model Details

AttnGAN AttnGan [60] is a text-to-image model that is introduced in 2017. It is based on an attention mechanism that allows the model to focus on different parts of the input text when generating an image. AttnGAN has been shown to generate high-quality images for a variety of datasets and has been widely used in a number of applications. However, the attention mechanism can be computationally expensive, and the model can be difficult to train.

X-LXMERT X-LXMERT [5] is an enhanced version of LXMERT [53]. It is introduced in 2020 and incorporates several training refinements. These refinements involve discretizing visual representations, utilizing uniform masking with a wide range of masking ratios, and aligning appropriate pre-training datasets to respective objectives.

minDALL-E MinDALL-E [27] is a fast, minimal port of Boris Dayma's DALL·E Mini (with mega weights). DALL·E Mini is an attempt to reproduce OpenAI's DALL-E [44] with a smaller architecture. DALL-E can generate high-quality new images from any text prompt. The checkpoint we use is DALL·E Mega, the latest version of DALL·E Mini.

VQ-Diffusion VQ-Diffusion [16] is a generative model that combines vector quantization (VQ) and diffusion-based models for image synthesis. VQ-Diffusion builds upon the framework of diffusion-based generative models, which involves simulating a stochastic process that gradually transforms a simple noise distribution into the target data distribution. In VQ-Diffusion, the image data is first quantized into discrete codes using a VQ algorithm, which maps each image patch to the nearest code in a codebook. This allows the model to represent complex data distributions with a compact set of discrete codes, rather than continuous probability densities.

Stable Diffusion Stable Diffusion is a pre-trained diffusion model for text-to-image generation. It is based on Latent Diffusion model (LDM) [46]. LDM is designed to learn the underlying structure of a dataset by mapping it to a lower-dimensional latent space. This latent space represents the data in which the relationships between different data points are more easily understood and analyzed, and reduces the amount of computational resources needed for training diffusion models. Specifically, we use three versions of Stable Diffusion, v1.1, v1.5, and v2.1. Each version is trained with a different number of steps and amount of data.

F. Prompt

For demonstration purposes, we show part of the prompt for question generation with GPT-3 in-context learning. The whole prompt will be released with our codes. The prompt contains instructions and several in-context examples. The examples cover all element categories.

```
Given an image description, generate
multiple-choice questions that verify if
the image description is correct.

First extract elements from the image
description. Then classify each element
into a category (object, human, animal,
food, activity, attribute, counting, color,
 material, spatial, location, shape, other).
 Finally, generate questions for each
element.

Description: A man posing for a selfie in a
 jacket and bow tie.
Entities: man, selfie, jacket, bow tie
Activities: posing
Colors:
Counting:
Other attributes:
Questions and answers are below:
About man (human):
Q: is this a man?
Choices: yes, no
A: yes
Q: who is posing for a selfie?
Choices: man, woman, boy, girl
A: man
About selfie (activity):
Q: is the man taking a selfie?
Choices: yes, no
A: yes
Q: what type of photo is the person taking?
Choices: selfie, landscape, sports,
portrait
A: selfie
About jacket (object):
Q: is the man wearing a jacket?
Choices: yes, no
A: yes
Q: what is the man wearing?
Choices:jacket, t-shirt, tuxedo, sweater
A: jacket
About bow tie (object):
Q: is the man wearing a bow tie?
Choices: yes, no
A: yes
Q: is the man wearing a bow tie or a neck
tie?
Choices: bow tie, neck tie, cravat, bolo
tie
A: bow tie
About posing (activity):
Q: is the man posing for the selfie?
Choices: yes, no
A: yes
Q: what is the man doing besides taking the
 selfie?
Choices: posing, waving, nothing, shaking
A: posing

Description: A horse and several cows feed
on hay.
Entities: horse, cows, hay
Activities: feed on
Colors:
Counting: several
Other attributes:
Questions and answers are below:
About horse (animal):
Q: is there a horse?
Choices: yes, no
A: yes
About cows (animal):
Q: are there cows?
Choices: yes, no
A: yes
About hay (object):
Q: is there hay?
Choices: yes, no
A: yes
Q: what is the horse and cows feeding on?
Choices: hay, grass, leaves, twigs
A: hay
About feed on (activity):
Q: are the horse and cows feeding on hay?
Choices: yes, no
A: yes
```

About several (counting):
Q: are there several cows?
Choices: yes, no
A: yes

Description: A red colored dog.
Entities: dog
Activities:
Colors: red
Counting:
Other attributes:
Questions and answers are below:
About dog (animal):
Q: is this a dog?
Choices: yes, no
A: yes
Q: what animal is in the picture?
Choices: dog, cat, bird, fish
A: dog
About red (color):
Q: is the dog red?
Choices: yes, no
A: yes
Q: what color is the dog?
Choices: red, black, white, yellow
A: red

Description: Here are motorcyclists parked outside a Polish gathering spot for women
Entities: motorcyclists, gathering spot, women
Activities: parked
Colors:
Counting:
Other attributes: outside, polish
Questions and answers are below:
About motorcyclists (human):
Q: are there motorcyclists?
Choices: yes, no
A: yes
About gathering spot (location):
Q: is this a gathering spot?
Choices: yes, no
A: yes
About women (human):
Q: are there women?
Choices: yes, no
A: yes
Q: who are in the gathering spot?
Choices: women, men, boys, girls
A: women
About parked (activity):
Q: are the motorcyclists parked?
Choices: yes, no
A: yes
About outside (spatial):
Q: have the motorcyclists parked outside the gathering spot?
Choices: yes, no
A: yes
Q: are the motorcyclists outside or inside of the gathering spot?
Choices: outside, inside, on the roof, in the basement
A: outside
About Polish (other):
Q: is this a Polish gathering spot?
Choices: yes, no
A: yes
Q: is this a Polish or a Chinese gathering spot?
Choices: Polish, American, Chinese, Japanese
A: Polish