TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering

Yushi Hu[1] Benlin Liu[1] Jungo Kasai[1] Yizhong Wang[1]
Mari Ostendorf[1] Ranjay Krishna[1,2] Noah A. Smith[1,2]
[1]University of Washington [2]Allen Institute for AI
https://tifa-benchmark.github.io/

Abstract

*Despite thousands of researchers, engineers, and artists actively working on improving text-to-image generation models, systems often fail to produce images that accurately align with the text inputs. We introduce TIFA (**T**ext-to-**I**mage **F**aithfulness evaluation with question **A**nswering), an automatic evaluation metric that measures the faithfulness of a generated image to its text input via visual question answering (VQA). Specifically, given a text input, we automatically generate several question-answer pairs using a language model. We calculate image faithfulness by checking whether existing VQA models can answer these questions using the generated image. TIFA is a* reference-free *metric that allows for fine-grained and interpretable evaluations of generated images. TIFA also has better correlations with human judgments than existing metrics. Based on this approach, we introduce TIFA v1.0, a benchmark consisting of 4K diverse text inputs and 25K questions across 12 categories (object, counting, etc.). We present a comprehensive evaluation of existing text-to-image models using TIFA v1.0 and highlight the limitations and challenges of current models. For instance, we find that current text-to-image models, despite doing well on color and material, still struggle in counting, spatial relations, and composing multiple objects. We hope our benchmark will help carefully measure the research progress in text-to-image synthesis and provide valuable insights for further research.*[1]

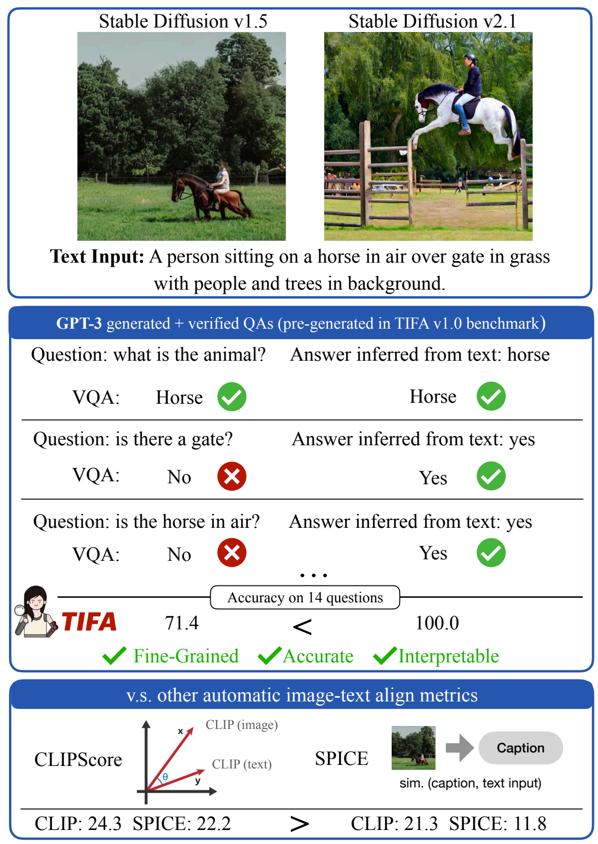

Figure 1. Illustration of how TIFA works, and comparison with the widely-used CLIPScore and SPICE metrics. Given the text input, TIFA uses GPT-3 to generate several question-answer pairs, and a QA model filters them (3 out of 14 questions for this text input are shown). TIFA measures whether VQA models can accurately answer these questions given the generated image. In this example, TIFA indicates that the image generated by Stable Diffusion v2.1 is better than that by v1.5, while CLIP and SPICE yield the opposite result. The text input is from the MSCOCO validation set.

1. Introduction

While we welcome artistic freedom when we commission art from artists, images produced by deep generative models [44, 46, 43, 47, 61] should conform closely to our requests. Despite the advances in generative models, it is still challenging for models to produce images faithful to users' intentions [40, 11, 30, 35, 36]. For example, current models often fail to compose multiple objects [40, 11, 35], bind attributes to the wrong objects [11], and struggle in generating visual text [36]. Today, there are efforts to address

[1]Correspondance to <Yushi Hu: yushihu@uw.edu>. All data and a pip-installable evaluation package are available on the project page.

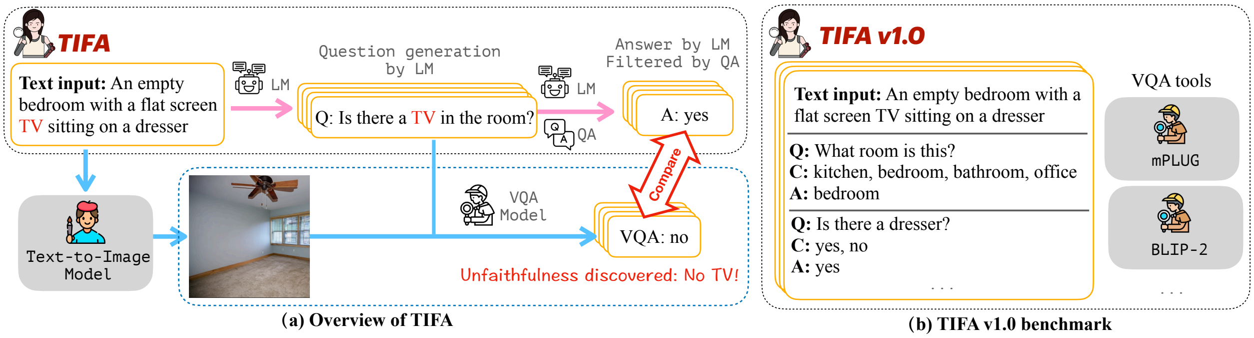

Figure 2. **(a) Overview of how TIFA evaluates the faithfulness of a synthesized image.** TIFA uses a language model (LM), a question-answering (QA) model, and a visual-question-answering (VQA) model. Given a text input, we generate several question-answer pairs with the LM and then filter them via the QA model. To evaluate the faithfulness of a synthesized image to the text input, a VQA model answers these visual questions using the image, and we check the answers for correctness. **(b) TIFA v1.0 benchmark.** While TIFA is applicable to any text prompt, to allow direct comparison across different studies, and for ease of use, we introduce the TIFA v1.0 benchmark, a repository of text inputs along with pre-generated question-answer tuples with answer choices. To evaluate a text-to-image model, a user first produces the images for the text inputs in TIFA v1.0 and then performs VQA with our provided tools on generated images to compute TIFA.

these challenges: researchers are imposing linguistic structure with diffusion guidance to produce images with multiple objects [11]; others are designing reward models trained using human feedback to better align generations with user intention [30]. However, progress is difficult to quantify without accurate and interpretable evaluation measures that explain when and how models struggle.

A critical bottleneck, therefore, is the lack of reliable automatic evaluation metrics for text-to-image generation faithfulness. One of the popular metrics is CLIPScore [17], which measures the cosine similarity between the CLIP embeddings [42] of the text input and the generated image. However, since CLIP is not effective at counting objects [42], or reasoning compositionally [37], CLIPScore is unreliable and often inaccurate. Another family of evaluation metrics uses image captions, in which an image captioning model first converts the image into text, and then the image caption is evaluated by comparing it against the text input. Unfortunately, using captioning models is insufficient since they might decide to ignore salient information in images or focus on other non-essential image regions [24]; for example, a captioning model might say that the images in Figure 1 are “a field of grass with trees in the background”. Moreover, evaluating text (caption) generation is inherently challenging [23, 26]. Another recent text-to-image evaluation is DALL-Eval [6], which employs object detection to determine if the objects in the texts are in the generated images. However, this approach only works on synthesized text and measures faithfulness along the limited axes of objects, counting, colors, and spatial relationships but misses activities, geolocation, weather, time, materials, shapes, sizes, and other potential categories we often ask about when we recall images from memory [29].

To address the above challenges, we introduce TIFA, a new metric to evaluate text-to-image generation faithfulness. Our approach is illustrated in Figure 2. Given a repository of text inputs, we automatically generate question-answer pairs for each text via a language model (here, GPT-3 [3]). A question-answering (QA) system (here, UnifiedQA [25]) is subsequently used to verify and filter these question-answer pairs. To evaluate a generated image, we use a visual-question-answering (VQA) system (here, mPLUG-large [31], BLIP-2 [32], etc.) to answer the questions given the generated image. We measure the image’s faithfulness to the text input as the accuracy of the answers generated by the VQA system. While the accuracy of TIFA is dependent on the accuracy of the VQA model, our experiments show that TIFA has much higher correlation with human judgments than CLIPScore (Spearman’s $\rho = 0.60$ vs. 0.33) and captioning-based approaches (Spearman’s $\rho = 0.60$ vs. 0.34). Additionally, since the LMs and VQA models will continue to improve, we hypothesize that TIFA will continue to be more reliable over time. Also, our metrics can automatically detect when elements are missing in the generation: in Figure 2, TIFA detects that the generated image does not contain a TV.

To promote the use of our new evaluation metric, we release TIFA v1.0, a large-scale text-to-image generation benchmark containing 4K diverse text inputs, sampled from the MSCOCO captions [34], DrawBench [47], PartiPrompts [61], and PaintSkill [6]. Each input comes with a pre-generated set of question-answer pairs, resulting in 25K questions covering 4.5K distinct elements. These questions have been automatically generated and pre-filtered using a question-answering model. This benchmark also comes with different VQA models [57, 28, 58, 32, 31, 21] that can be used to evaluate generative models and can be easily extended to use future VQA models when they become available.

We conduct a comprehensive evaluation of current text-to-image models using TIFA v1.0. Thanks to TIFA's ability to detect fine-grained unfaithfulness in images, we find that current state-of-the-art models are good at rendering common objects, animals, and colors, but still struggle in composing multiple objects, reasoning about spatial relations, and binding the correct activity for each entity. In addition, our ablation experiments show that TIFA is robust to different VQA models. Future researchers can use TIFA v1.0 to compare their text-to-image models' faithfulness across different studies. Also, future generative models may focus on addressing the weaknesses of current models that TIFA discovered. In addition, with TIFA, users can customize evaluations with their own text inputs and questions [10]; for example, a future TIFA benchmark could focus on counting or scene text.

2. Related Work

We compare TIFA to other image and language generation evaluation metrics.

Prior image generation evaluation Prior work usually compares image generation models via pairwise comparison by humans. How to design automatic evaluation metrics to approximate human assessment of the quality of machine-generated images has always been a major challenge in computer vision. There are two aspects to evaluate, namely image quality and image-text alignment. **Inception Score** [48] and **FID** [18] are the most widely adopted metrics for image quality. They compare the features of the generated images and gold images extracted from a pre-trained Inception-V3 model [52] to evaluate the fidelity and diversity of generated images. However, they rely on ground-truth images and are based on a classification model, which makes them not suitable for complex datasets [12]. For image-text alignment, prior metrics are mainly based on CLIP [42], captioning, and object detection models. **CLIPScore** [17] and **CLIP-R** [39] are based on the cosine similarity of image and text CLIP [42] embeddings. [6, 19, 20] first convert the images using a captioning model, and then compare the image caption with the text using metrics like CIDEr [55] and SPICE [1]. **SOA** [19] and **DALL-Eval** [6] employ object detection models to determine if objects, attributes, and relations in the text input are in the generated image. However, this approach only works on synthesized text inputs and measures faithfulness on limited axes (object, counting, color, spatial relation), missing elements like material, shape, activities, and context. In contrast, thanks to the flexibility of questions, TIFA works on any text inputs and evaluates faithfulness across a broad spectrum of dimensions.

Summarization evaluation in NLP TIFA is inspired by the summarization evaluation methods based on question answering (QA) [56, 50]. Given a summary, a language model generates a set of questions about the text. A QA model checks if the same answer can be inferred from the text and the summary. These QA-based metrics have much higher correlations with human judgments on the factual consistency of summarization than other automatic metrics [56, 50]. TIFA can be seen as treating the text input for the text-to-image model as a summary of the generated image.

3. The TIFA Metric

We introduce a framework for automatically estimating the faithfulness of an image to its text prompt. Given a text input T, we aim to measure the faithfulness of the generated image I. An overview of our metric is illustrated in Figure 2. From T, we generate N multiple-choice question-answer tuples $\{Q_i, C_i, A_i\}_{i=1}^N$, in which Q_i is a question, C_i is a set of answer choices, and $A_i \in C_i$ is the gold answer. The answer A_i can be inferred given T, Q_i, and C_i. Next, for each question Q_i, we use a VQA model to produce an answer $A_i^{\text{VQA}} = \max_{a \in C_i} p(a \mid I, Q_i)$. We define the faithfulness between the text T and image I as the VQA accuracy:

$$\textit{faithfulness}(T, I) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[A_i^{\text{VQA}} = A_i] \tag{1}$$

The range of our faithfulness score is $[0, 1]$. It is maximized when we have a performant VQA model, and the image I accurately covers the information in the text T so that for any question Q, which can be answered given T can also be answered given I. Several key design decisions will be addressed in later sections: how to generate questions (§3.1), how to control the question quality (§3.2), and how to answer those questions (§3.3). Finally, we give a step-by-step qualitative example of TIFA in Figure 4.

3.1. Question-Answer Generation

Our main challenge is to generate diverse questions that cover all elements of the text input evenly. We also simplify the question-generation pipeline into a single GPT-3 [3] completion, so that TIFA can exploit the power of recent language models (LM) and work with updated black-box LMs (e.g., ChatGPT) in the future.

Inspired by prior work [4], given a text prompt T, we generate the question-answer tuples $\{Q_i, C_i, A_i\}_{i=1}^N$ via the pipeline illustrated in Figure 3. Different from prior work, which relies on multiple components, our pipeline is completed by a single inference run via in-context learning with GPT-3 [3, 59, 22, 41, 51], thereby avoiding the need for intermediate human annotations. We annotate 15 examples and use them as in-context examples for GPT-3 to follow. Here we take the text *"A photo of three dogs."* as an example. Each in-context example contains the following steps:

Element extraction Given text prompt T, GPT-3 will first extract all elements $\{v_i\}_{i=1}^m$ following prior work [4] (for the in-context examples, we perform element extraction manually). The elements include noun phrases (including named

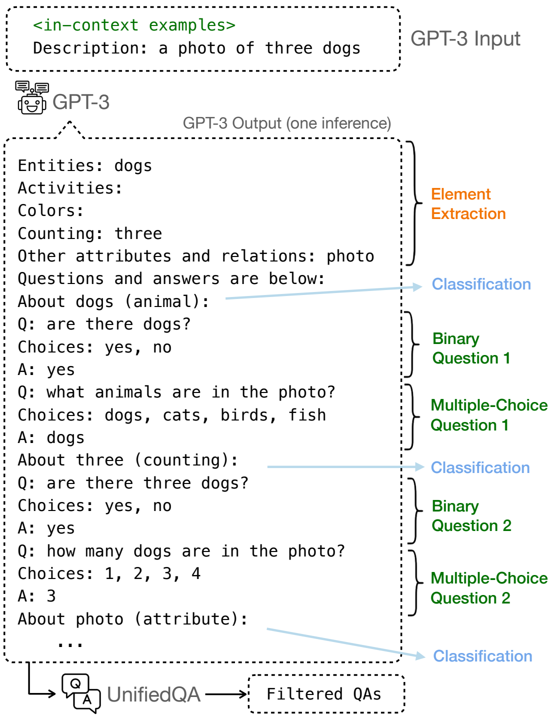

Figure 3. Our question-answer pair generation pipeline. The whole pipeline can be executed via a single inference of GPT-3 via in-context learning. Given the text prompt, GPT-3 first extracts the elements and then generates two questions for each element. The GPT-3 output is then parsed and filtered by UnifiedQA.

entities), verbs, adjectives, adverbs, and parse tree spans with no more than 3 words altogether. For the above example, the elements are *photo, three, dogs*.

Element category classification For each element v_i, following [29], we classify the elements into one of the following 12 categories: *object, activity, animal, food, counting, color, material, spatial, location, shape, attribute*, and other. As shown in Figure 3, text generated from GPT-3 contains the category corresponding to each question.For example, "three" is "counting", and "dogs" is classified as "animal." This step allows a detailed analysis of the text-to-image model's ability in each category.

Question generation conditioned on elements For each element v_i, we generate two questions. The first is a question that should be answered "yes" for a faithful generated image, and the second question has v_i as its answer. For example, two questions are generated for the element *"three"*. The first is "are there three dogs?", and the choices are {yes, no}. The second is "how many dogs are there?", and the choices are {1,2,3,4}. These two types of questions make our evaluations diverse and robust to surface-level differences.

Completing above steps by prompting GPT-3 once As mentioned earlier, for each text T, the whole pipeline can be completed by one GPT-3 inference. We annotated 15 in-context examples that cover all types of questions. The full prompt is in the Appendix. The prompt format is shown in Figure 3. Our in-context examples follow the same format, and identical examples are used for all text inputs, leading to a fixed and limited amount of human annotation cost. We use *code-davinci-002* engine for question generation, and the decoding temperature is 0.

3.2. Question Filtering

To ensure the quality of generated images, we use UnifiedQA [25] to verify the GPT-3 generated question-answer pairs and filter out the ones that GPT-3 and UnifiedQA do not agree on. UnifiedQA[2] is a state-of-the-art multi-task question-answering model that can answer both multiple-choice and free-form questions. Denote the UnifiedQA model as QA. Given the text T, question Q_i, choices C_i, and answer A_i, Let $A_i^f = QA(T, Q_i)$ be the free-form answer, and $A_i^{mc} = QA(T, Q_i, C_i)$ be the multiple-choice answer. We keep the question if $A_i = A_i^{mc}$ and the word-level F_1 score between A_i^f and A_i is greater than 0.7. We conduct a human evaluation on 1000 filtered question-answer pairs. Only 7 are considered not reasonable (e.g., generated choices do not include a correct answer). Details are in Appendix C.

Recommended VQA model Based on considerations over the accuracy, correlation with human judgments, and run time, we currently suggest using **mPLUG-large** as the VQA model for TIFA. Analysis is given in Section 5.4. Like the LM and QA components, the VQA component can be updated in the future as the technology improves.

3.3. VQA Models

Since our questions contain a diverse set of visual elements (e.g., activity, art style), we use open-domain pre-trained vision-language models as our VQA model (rather than closed-class classification models fine-tuned on VQAv2 [14]). We provide tools to easily perform VQA on arbitrary images and questions, based on 5 state-of-the-art VQA models trained with distinct data and strategies.

Vision-language models The general pre-trained vision-language model are **GIT-large** [57], **VILT-B/32** [28], **OFA-large** [58], and **mPLUG-large** [31]. These models are pre-trained on a large amount of image-text pairs, and down-stream image-to-text tasks like image captioning and visual question answering. Notice that these models have not been trained to answer multiple-choice questions. For each question, we first decode the free-form answer and then choose the choice that has the highest similarity with the decoded answer, measured by SBERT [45]. Another model we use is **BLIP-2 FlanT5-XL** [32], in which a VIT [8] is connected with a frozen FlanT5 [7] via a lightweight transformer. This

[2]Model checkpoint we use: https://huggingface.co/allenai/unifiedqa-v2-t5-large-1363200.

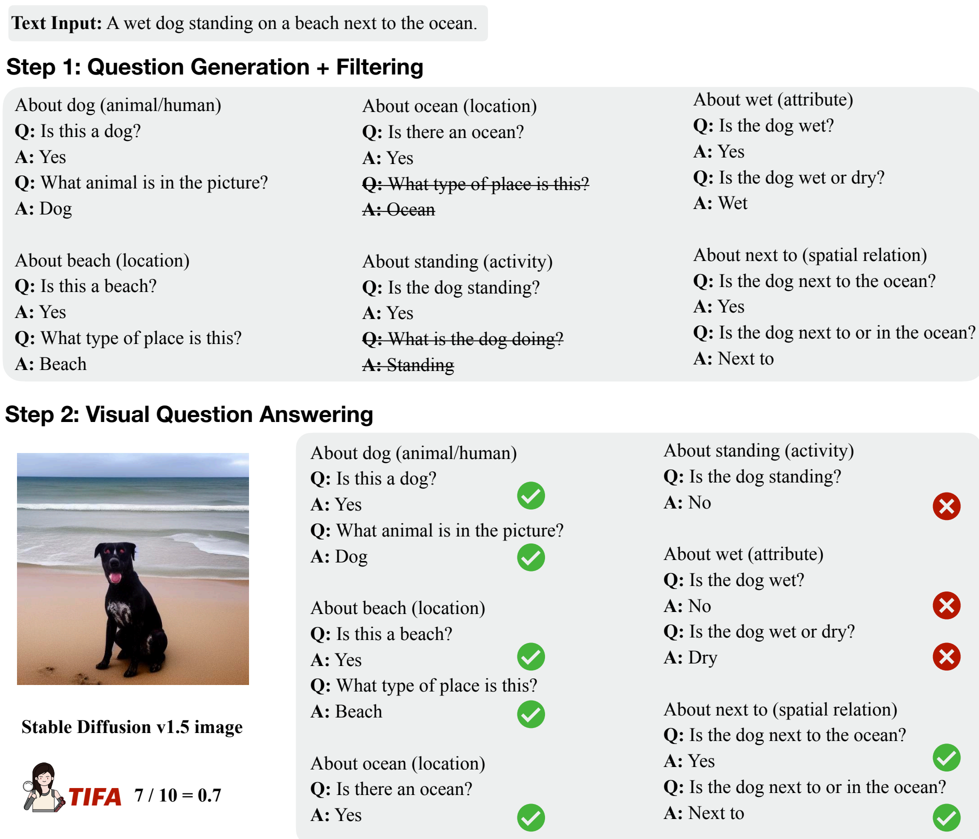

Figure 4. Step-by-step qualitative example of TIFA metric. Given a text input, we first generate question-answer pairs and filter them. We ~~strikethrough~~ the questions filtered out by UnifiedQA. Then we run VQA models on the generated image to get the TIFA score.

model allows for performing multiple-choice VQA directly due to the flexibility of the LM.

4. TIFA v1.0: Benchmark for Text-to-Image Generation Faithfulness

In this section, we introduce TIFA v1.0, a text-to-image generation faithfulness benchmark based on the evaluation method discussed in Section 3. The benchmark consists of 4,081 diverse text inputs paired with 25,829 question-answer pairs. Each question is classified into one of the categories discussed in Section 3.1. The benchmark also comes with Python pip-installable APIs to perform VQA with various state-of-the-art VQA models on arbitrary visual questions. The overall TIFA for each text-to-image model is computed by averaging TIFA scores of images generated from each text input in the benchmark.

4.1. Text Collections

We collect 4,081 text inputs to benchmark text-to-image models' generation ability on diverse tasks. 2,000 text inputs are image captions from **COCO** validation set [34]. These captions have corresponding gold images. Since text-to-image models are often used to create abstract art, we also collect 2,081 text inputs from previous works that do not correspond to any real image. All text inputs we use contain $\geq$ 3 words. We include 161 from **DrawBench** used in Imagen [47] (texts that are categorized as "misspellings" and "rare words" are removed); 1420 from **PartiPrompt** used in Parti [61] (texts in category "abstract" are removed); and 500 texts from **PaintSkill** used in DALL-Eval [6].

Table 1. Statistics of TIFA v1.0.

Statistics	
# of prompts	4,081
- # of COCO captions	2,000
- # of DrawBench, PartiPrompt, PaintSkill prompts	2,081
# of questions	25,829
- # of binary questions	17,226
- # of multiple-choice questions	8,603
avg. # of questions per prompt	6.3
avg. # of words per prompt	10.5
avg. # of elements per prompt	4.3

4.2. Statistics and Diversity

Table 1 shows the basic statistics of the TIFA v1.0 benchmark. We demonstrate TIFA v1.0's diversity in Figure 5.

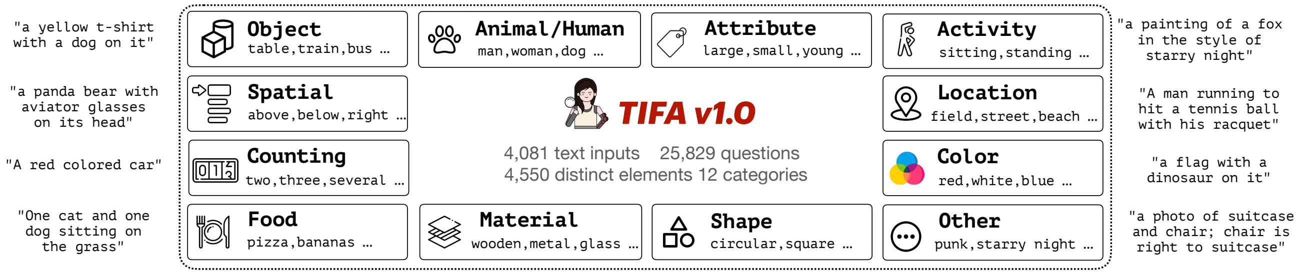

Figure 5. Statistics and diversity of TIFA v1.0. The text inputs contain elements from 12 categories (e.g., object, spatial, and counting). We show the most common elements from each category. In addtion, we also show some example text inputs on the sides.

TIFA v1.0 contains questions about 4,550 distinct elements, which are categorized into 12 categories. The number of times each type of element occurs in the text input is object (7,854), animal/human (3,501), attribute (3,399), activity (2,851), spatial (2,265), location (1,840), color (1,743), counting (986), food (911), material (209), shape (69), and other (201). The "other" category includes notions often used in abstract art, such as “starry night” and “steampunk.” The accuracy of VQA on a particular genre measures the text-to-image model’s ability in the corresponding aspect. Please refer to Appendix D for more details on TIFA v1.0.

4.3. Finetuned Open-Source Language Model for Question Generation

For TIFA v1.0, we use GPT-3 to generate the questions. While benchmarking with TIFA v1.0 is a deterministic process, using TIFA to create a new benchmark might not be deterministic as the underlying question generator (GPT-3 in our case) might change privately. To promote deterministic benchmark generation, we fine-tune and release a LLaMA 2 (7B) [54] model that parses the captions and generates questions for arbitrary texts, using TIFA v1.0 questions as training examples. [3]

5. Experiments

In this section, we first show that TIFA has substantially higher correlations with human judgments than prior metrics on text-to-image faithfulness (§5.1). Then we present a comprehensive evaluation of existing text-to-image models using TIFA v1.0, highlighting the challenges of current text-to-image models (§5.3). Finally, we conduct an analysis of TIFA’s robustness against different VQA models (§5.4). For all experiments, we use **mPLUG** as the VQA model for TIFA unless stated otherwise. The models we evaluate include AttnGAN [60], X-LXMERT [5], VQ-Diffusion [16], minDALL-E [27], and Stable Diffusion v1.1, v1.5, and v2.1 [46]. Details are in Appendix E.

[3]LLaMA 2 question generation model checkpoint: https://huggingface.co/tifa-benchmark/llama2_tifa_question_generation

5.1. Correlation with Human Judgements

To compare TIFA with prior evaluation metrics, we first conduct human evaluations of the text-to-image models on the 1-5 Likert scale on text-to-image faithfulness. Then we compare TIFA with other metrics based on their correlation with human judgments.

Likert scale on text-to-image faithfulness Annotators are asked to answer on a scale of 1 (worst) to 5 (best) to the question *“Does the image match the text?"*. The detailed annotation guidelines are in Appendix C. Annotators are asked to focus on text-to-image faithfulness rather than image quality. The Likert scale should be based on how many elements in the text prompt are missed or misrepresented in the image. Objects are more important than attributes, relations, and activities. If an object is missed in the image, then all related attributes, activities, relations, etc. are also considered lost. An example is given in Figure 6.

We collect annotations of 800 generated images on 160 text inputs from TIFA v1.0. For each prompt, we sample an image from the 5 most recent generative models we evaluated, i.e., minDALL-E, VQ-Diffusion, Stable Diffusion v1.1, v1.5, and v2.1. We collect 2 annotations per image and average over the scores as the single “faithfulness" score. The inter-annotator agreement measured by Krippendorf’s α is 0.67, indicating “substantial" agreement.

Figure 6. Illustration of our Likert scale annotation guideline. Annotators are asked to give a score of 1 to 5 based on how many elements in the text prompt are missed or misrepresented in the image. The missed elements are underlined.

Baselines We compare our evaluation with two families of reference-free metrics on text-image match introduced in

Section 2. The first is the **caption-based method**. We use the state-of-the-art **BLIP-2 FlanT5-XL** [32] as the captioning model. The second approach is **CLIPScore** [17, 42]. We use CLIP (VIT-B/32) [42] to compute the score.

Table 2. Correlations between each evaluation metric and human judgment on text-to-image faithfulness, measured by Spearman's ρ and Kendall's τ.

	Spearman's ρ	Kendall's τ
Caption-Based		
BLEU-4	18.3	18.8
ROUGE-L	32.9	24.5
METEOR	34.0	27.4
SPICE	32.8	23.2
CLIPScore	33.2	23.1
Ours		
TIFA (VILT)	49.3	38.2
TIFA (OFA)	49.6	37.2
TIFA (GIT)	54.5	42.6
TIFA (BLIP-2)	55.9	43.6
TIFA (mPLUG)	**59.7**	**47.2**

TIFA has a much higher correlation with human judgments than prior metrics. The correlations between each evaluation metric and human judgment are shown in Table 2. For caption-based evaluations, we use metrics BLEU-4 [38], ROUGE-L [33], METEOR [2], and SPICE [1]. TIFA has higher correlations with human judgments than all previous evaluation metrics on all VQA models. TIFA (mPLUG) yields the highest correlation with human judgments among all VQA models.

5.2. Benchmarking Text-to-Image Models

Figure 7 shows the average TIFA score text-to-image models get on TIFA v1.0. The detailed scores with each VQA model on each element type are provided in Appendix B. We can see a clear trend of how text-to-image models evolve over time. There is a jump in TIFA score after DALL-E [44] is released, from about 60% to 75%. Qualitative examples of our evaluation metric are in Appendix A.

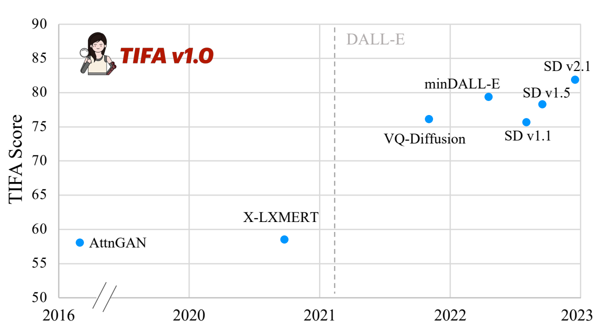

Figure 7. Average TIFA score of text-to-image models on the TIFA v1.0 benchmark. The horizontal axis shows their release dates.

5.3. Findings on Current Text-to-Image Models

Figure 8 shows accuracy on each type of question in TIFA v1.0 for Stable Diffusion v1.1, v1.5, and v2.1. The score on each type reflects the text-to-image models' faithfulness in each type of visual element. To the best of our knowledge, TIFA is the only automatic evaluation method that can provide such a detailed fine-grained analysis of image generation. We separate the scores on COCO captions and other text inputs. For COCO captions, we also include the accuracy on the ground-truth images for reference. We summarize our findings in the following paragraphs.

Generating images from captions vs. free-form text From Figure 8, we can see that VQA accuracy is higher on the COCO captions than on other text inputs. The reason is that COCO captions correspond to real images, while other text inputs may correspond to compositions that cannot be found in real-world photos (e.g. "a blue apple").

What elements are text-to-image models struggling with? Based on the scores of each category in Figure 8, we can see that Stable Diffusion models are performing well on material, animal/human, color, and location in terms of text-to-image faithfulness. However, they yield low accuracy on questions involving **shapes**, **counting**, and **spatial relations**. "Other" mainly contains **abstract art notions**, and models are also struggling with them. There is also a big gap between the synthesized images and real images on the COCO captions. Future work can explore various directions (e.g., training data/loss and model architecture) to improve text-to-image models' faithfulness in these aspects.

Why are ground-truth images not getting perfect scores? Ground-truth images in COCO do not get perfect scores because 1) the COCO captions contain a substantial amount of noise from crowd workers [24] and 2) VQA models are not perfect. Real images have higher accuracy in all categories except material, color and location, where differences are small. It is left to future work to determine whether this is simply due to noise or it is an area where assessment can be improved.

Stable Diffusion is evolving. We can see the consistent trend that Stable Diffusion models are improving in their later versions in most of the element categories. The exceptions are "shape" for both prompt sources, "other" and "food" for the free-form text inputs without gold images.

Composing multiple objects is challenging. Figure 9 shows how the number of entities (objects, animals/humans, food) in the text input affects the average TIFA score. When there are more than 5 entities, The TIFA score starts to drop rapidly for all text-to-image models, consistent with similar findings in other vision-language evaluations [15, 13]. For reference, we also add the real images in COCO in this figure.

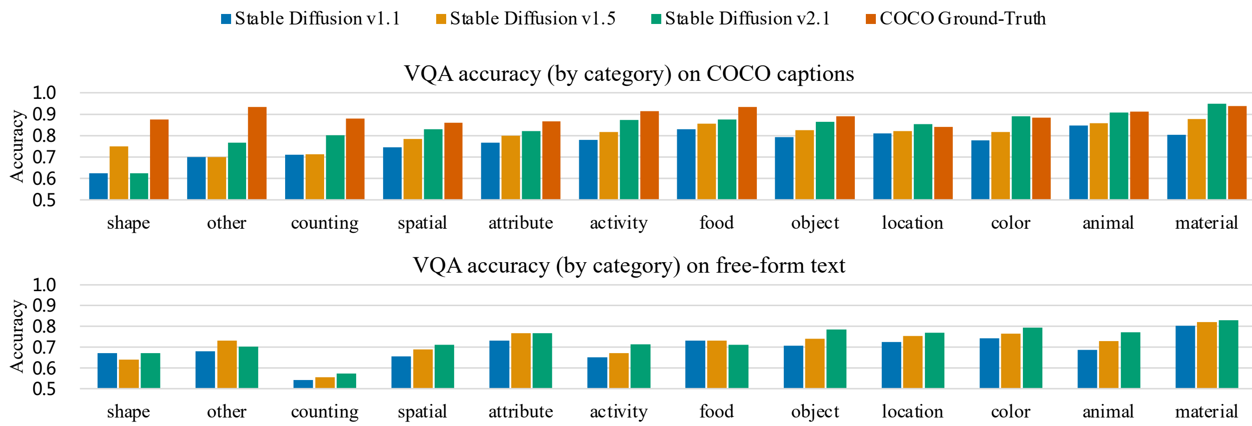

Figure 8. Accuracy on each type of question in the TIFA v1.0 benchmark. The text-to-image models are Stable Diffusion v1.1, v1.5, and v2.1. We order the categories by the average score Stable Diffusion v2.1 gets on corresponding questions. For COCO captions, we also include the accuracy of the ground-truth images for reference.

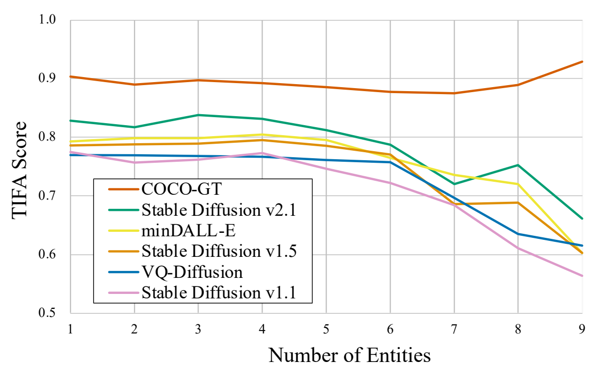

Figure 9. TIFA vs. numbers of entities (objects, animals/humans, and food) in the text input. The accuracy starts to drop when more than 5 entities are added to the text, showing that compositionality is hard for text-to-image models. Meanwhile, TIFA scores for COCO ground-truth (GT) images remain consistent.

The TIFA score on real images is rather consistent and does not change as the number of entities increases. This quantitatively shows that composing multiple objects is challenging for current text-to-image models. One possible reason is that the CLIP text embedding, which is used to train Stable Diffusion, lacks compositionality, as investigated in [37].

5.4. Analysis of VQA Models

One major concern of TIFA is that VQA models can introduce some errors. Table 2 shows that TIFA has a much higher correlation with human judgment than the previous metrics, regardless of the choice of the VQA models; here we conduct a more detailed analysis.

Sensitivity of TIFA to VQA models Figure 10 shows several recent text-to-image models' TIFA scores on the COCO captions in TIFA v1.0, measured by different VQA models. We also include the TIFA scores on the ground-truth COCO

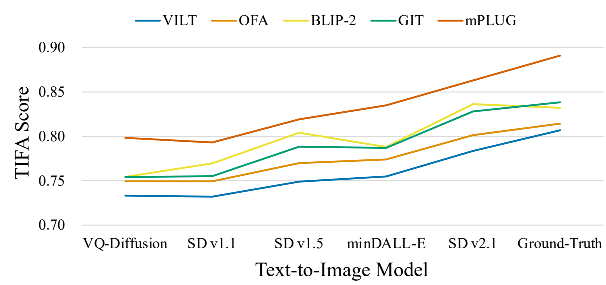

Figure 10. Several text-to-image models' TIFA score on COCO captions, measured by different VQA models. We also include the accuracy of ground-truth images for reference.

images for reference. TIFA scores computed by different VQA models show a similar trend on these text-to-image models. Also, the ground-truth images get the highest TIFA score. We also computed Spearman's ρ of TIFA scores given by different VQA models. The pairwise correlation between all VQA models is greater than 0.6.

Humans performing VQA To conduct further analysis on the VQA models, we ask annotators to answer the multiple-choice visual questions in TIFA v1.0. These annotations help us evaluate the accuracy of each VQA model. For multiple-choice questions, we add the option "None of the above" for human evaluation. Annotation guidelines are in Appendix C.

We collect annotations of 1029 questions on 126 images. Each question is answered by two annotators. The inter-annotator agreement measured by Krippendorf's α is 0.88. A third annotator is involved if two annotators disagree, and the final answer is chosen by the majority vote.

Which VQA model should we use?

Table 3 reports the accuracy of each VQA model and the correlation between TIFA scores calculated by VQA model answers and human answers. We observe that higher model performance is directly related to the TIFA score's

Table 3. Comparison of VQA models. The first row is the VQA accuracy, using the human VQA answers as reference. The second row is Spearman's correlation between TIFA scores calculated by each VQA model and the human VQA.

	VILT	OFA	GIT	BLIP-2	mPLUG
VQA Acc.	76.1	77.1	79.1	81.0	84.5
TIFA Corr.	60.9	63.7	72.5	75.6	76.8

correlation with human judgments. **mPLUG** has the highest accuracy.

Another important factor to consider is the runtime. We measure the inference speed of each VQA model on NVIDIA A40 GPU with batch size 1 over the Stable Diffusion v2.1 images (768 × 768 pixels). For one question, VILT takes 0.08s on average; OFA, GIT, and mPLUG all take about 0.25s; BLIP-2 takes 0.73s. Based on the above results, we choose **mPLUG** as the default VQA model for TIFA v1.0 because it is the most accurate while being reasonably fast.

Separation of Text-to-Image Errors and VQA Errors Suppose an image gets a wrong answer given a visual question. Then the image generation or the VQA model might have made an error. Based on the human VQA results, we separate these two kinds of errors in Figure A. If human VQA gives the wrong answer, then we suspect the generated image has an error. Otherwise, the image is correct but the VQA model is making an error. Figure A shows that the majority of errors are made by the text-to-image models. For mPLUG, less than 25% errors are due to the VQA model. This suggests that the TIFA framework is a viable evaluation method despite its inherent challenges.

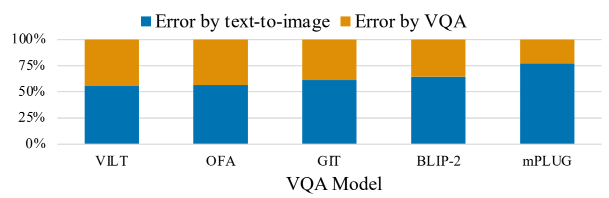

Figure 11. Source of the error when VQA gets the wrong answer.

6. Discussion

Why does TIFA work better than CLIPScore? Our experimental results and human evaluations in Section 5 show that TIFA is more accurate than prior metrics (CLIP [42] and captioning-based approaches) for evaluating text-to-image faithfulness. We hypothesize that the major challenge of these prior metrics is that they summarize the image outputs and text inputs into a single representation (embedding/caption). In contrast, TIFA exploits the power of the language models to decompose the text input into fine-grained probes, which allows VQA to capture more nuanced aspects of the text input and the generated image.

How do current VQA models perform on TIFA? One limitation is that TIFA requires VQA models to work reasonably well, which is true given the current models and TIFA v1.0, as shown in Section 5. Nevertheless, the assumption might not hold for current models in domains like anime and abstract art. TIFA is a modularized evaluation framework. The VQA models used within the framework can be updated as stronger VQA models become available in the future. For instance, we plan to incorporate GPT-4 once its image API is made public since it is likely to improve TIFA. Another possible solution is to ensemble multiple image understanding models. For example, one may employ expert models on art concepts. We leave this for future work.

Other limitations. Another limitation of TIFA is its runtime. Answering multiple visual questions is slower than one CLIP inference. In the scenario described in Section 5.4, mPLUG takes 1.6s to evaluate one image (without batching). Also, our question generation pipeline needs one inference on a modern language model for each text input. The run time is not a critical issue for benchmarking purposes, but may not be computationally feasible for the kind of large-scale data filtering done, for example, in LAION-5B [49]. Nevertheless, we would like to point out that our evaluation is much faster than the image generation process of diffusion models. Thus, we believe it is feasible to perform reranking and reinforcement learning with TIFA on diffusion models.

7. Conclusions

We present TIFA, a new automatic text-to-image faithfulness evaluation metric using VQA. Compared with prior metrics, TIFA is fine-grained, interpretable, and better aligned with human judgments. Based on this metric, we introduce the TIFA v1.0, a large-scale text-to-image benchmark containing 4K prompts and 25K questions. We conduct a comprehensive study of current text-to-image models using TIFA v1.0 and highlight the limitations of current generative models. We quantitatively show that current image generation models still struggle in counting, spatial relations, and composing multiple objects. Finally, we conduct extensive analysis and human evaluation, demonstrating that TIFA is robust to different VQA models. We hope TIFA will help evaluate future work on image generation and become increasingly sophisticated as it is upgraded with new LM, QA, and VQA components.

8. Acknowledgements

We thank Zirong Ye, Cheng-Yu Hsieh, Oscar Michel, Enhao Zhang, Jiafei Duan, Weijia Shi, Jieyu Zhang, the TIAL group, the ARK group, and the GRAIL lab at UW for their helpful feedback on this work. We would also thank all anonymous reviewers for providing insightful and constructive comments on our work.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 3, 7

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3, 13

[4] Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *North American Chapter of the Association for Computational Linguistics*, 2022. 3

[5] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-lxmert: Paint, caption and answer questions with multi-modal transformers. *ArXiv*, abs/2009.11278, 2020. 6, 14, 15

[6] Jaemin Cho, Abhaysinh Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *ArXiv*, abs/2202.04053, 2022. 2, 3, 5

[7] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022. 4

[8] 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. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4

[9] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2020. 14

[10] Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 3

[11] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, P. Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *ArXiv*, abs/2212.05032, 2022. 1, 2

[12] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas R. Dengel. Adversarial text-to-image synthesis: A review. *Neural networks : the official journal of the International Neural Network Society*, 144:187–209, 2021. 3

[13] Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5046–5055, 2022. 7

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[15] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 7

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10686–10696, 2021. 6, 14, 15

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Joseph Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing*, 2021. 2, 3, 7

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 3

[19] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:1552–1565, 2019. 3

[20] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018. 3

[21] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022. 2

[22] Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. In-context learning for few-shot dialogue state tracking. *arXiv preprint arXiv:2203.08568*, 2022. 3

[23] Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *North American Chapter of the Association for Computational Linguistics*, 2022. 2

[24] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Daniel Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. Transparent human evaluation for image captioning.

In *North American Chapter of the Association for Computational Linguistics*, 2022. 2, 7

[25] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings*, 2020. 2, 4, 13

[26] Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel Weld. GENIE: Toward reproducible and standardized human evaluation for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 2

[27] Saehoon Kim, Sanghun Cho, Chiheon Kim, Doyup Lee, and Woonhyuk Baek. mindall-e on conceptual captions, 2021. 6, 14, 15

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2, 4

[29] Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Information maximizing visual question generation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2008–2018, 2019. 2, 4

[30] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 1, 2

[31] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng da Cao, Ji Zhang, Songfang Huang, Feiran Huang, Jingren Zhou, and Luo Si. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In *Conference on Empirical Methods in Natural Language Processing*, 2022. 2, 4, 13

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 2, 4, 7

[33] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 7

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5, 13, 15

[35] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. *ArXiv*, abs/2206.01714, 2022. 1

[36] Rosanne Liu, Daniel H Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, R. J. Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. *ArXiv*, abs/2212.10562, 2022. 1

[37] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *ArXiv*, abs/2212.07796, 2022. 2, 8

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 7

[39] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 3

[40] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*, 2022. 1

[41] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models, 2022. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 7, 9, 14

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 1, 7, 15

[45] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084, 2019. 4

[46] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1, 6, 14, 16

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 1, 2, 5, 13, 15

[48] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *ArXiv*, abs/1606.03498, 2016. 3

[49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next gener-

ation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 9

[50] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 3

[51] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2015. 3

[53] Hao Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *ArXiv*, abs/1908.07490, 2019. 15

[54] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[55] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. 3

[56] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 3

[57] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2, 4

[58] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022. 2, 4

[59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3

[60] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2017. 6, 14, 15

[61] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv*, abs/2206.10789, 2022. 1, 2, 5