# Offline Evaluation of Set-Based Text-to-Image Generation

Negar Arabzadeh*
University of Waterloo
Waterloo, Canada
narabzad@uwaterloo.ca

Fernando Diaz
Google
Pittsburgh, USA
diazf@acm.org

Junfeng He
Google Inc.
Mountain View, USA
junfenghe@google.com

## Abstract

Text-to-Image (TTI) systems often support people during ideation, the early stages of a creative process when exposure to a broad set of relevant images can help explore the design space. Since ideation is an important subclass of TTI tasks, understanding how to quantitatively evaluate TTI systems according to how well they support ideation is crucial to promoting research and development for these users. However, existing evaluation metrics for TTI remain focused on distributional similarity metrics like Fréchet Inception Distance (FID). We take an alternative approach and, based on established methods from ranking evaluation, develop TTI evaluation metrics with explicit models of how users browse and interact with sets of spatially arranged generated images. Our proposed offline evaluation metrics for TTI not only capture how relevant generated images are with respect to the user's ideation need but also take into consideration the diversity and arrangement of the set of generated images. We analyze our proposed family of TTI metrics using human studies on image grids generated by three different TTI systems based on subsets of the widely used benchmarks such as MS-COCO captions and Localized Narratives as well as prompts used in naturalistic settings. Our results demonstrate that grounding metrics in how people use systems is an important and understudied area of benchmark design.

## CCS Concepts

• **Information systems** → **Information systems applications**; **Information retrieval**; **Evaluation of retrieval results**; • **Computing methodologies** → **Computer vision**.

## Keywords

Text to Image Generation, Offline Evaluation, User modelling

---

*work done while at Google

---

## 1 Introduction

Generative models and specifically text-to-image (TTI) models are being increasingly used in creative tasks, especially through ideation, the process of generating a list of design candidates in the beginning of a creative design task [28, 41, 44, 48, 96]. In this paper, we study TTI systems, a class of models particularly well-suited for ideation in the visual arts [47], architecture [64], and crafting [88]. In order to robustly benchmark TTI systems, researchers need to develop both datasets (i.e., prompts and relevance information such as a target image) [68, 78, 98] as well as *evaluation metrics* [32, 62]. These metrics can either be based on human rater preferences between a pair of system outputs (e.g., through a side-by-side comparison for a fixed prompt) or on offline metrics based on the automatic comparison of model output to an example system target.

While there have been advances in the critique and mitigation of datasets involved in benchmarking TTI systems [76, 84], evaluation metrics remain relatively immature and deserving of more attention [62]. Evaluation methods based on manual side-by-side comparisons of system output [2–4, 6], although effective at detecting differences between systems, suffer from multiple inefficiencies. On the one hand, this method incurs time costs arising from recruiting and training participants as well as conducting the assessment alone; when operating in rapidly advancing environments, these time costs can limit the space of models evaluated (and developed). On the other hand, because each pair of systems needs to be manually compared, the evaluation scales quadratically in the number of systems [29]; as a result, the financial cost of model evaluation may be prohibitive. Offline or automated evaluation metrics address these scalability issues by using information such as target image data to algorithmically score and compare systems [37, 79]. The most popular approach is to use the Fréchet Inception Distance (FID) [37], a method that compares the distribution of generated images over a workload of prompts to the distribution of example or target images over the same workload of prompts. FID, while mathematically convenient, is deficient in several ways [10, 25]. First, FID compares population-level statistics and, therefore, a system can generate a set of images comparable to the target set while individually being poor responses to the prompt. Second, because FID is a population-level metric, the benefits of per-prompt error analysis disappear. This includes evaluating for sub-population metrics such as fairness. Moreover, when comparing a pair of systems, because FID measures population level differences, evaluation cannot leverage matched statistical tests, known to improve sensitivity. Finally, FID may inaccurately measure TTI effectiveness because it is computed using a collection of generated images within and across prompts without paying attention to how they may be used or interacted with by users. This risks metric divergence [58], when there is a disconnection between an evaluation metric and the construct we are interested in measuring, a more general version of

the value alignment problem [93]. FID does not explicitly model user behavior or task, even though we know users engage with TTI systems in the context of ideation through a grid interface. As a result, evaluation may be brittle, especially as systems become effective enough to be operating in a region of the performance landscape where FID cannot effectively detect differences (cf. similar phenomenon in search evaluation [89]).

Seeing the need for reusable offline evaluation as a necessary benchmarking tool, we develop novel TTI evaluation metrics based on explicit models of how users engage with system decisions. This responds to a recent call to abandon metrics like FID that do not model human behavior and perception [62]. We leverage theories from the ideation literature to understand what people need during ideation [35, 45, 61, 80] and techniques from Information Retrieval (IR) to formally model these concepts [17]. In particular, we adapt two well-studied IR metrics, expected reciprocal rank [18] and rank-biased precision [57], to account for grid layouts, generated images, and ideation intent. We note that to leverage the soft relevance labels based on similarities to example images, in this study, we assume that there is always one or more ground truth example images known to be relevant to the given prompt.

We conduct experiments on comparing our proposed set of metrics with traditional ones, demonstrating that (i) modeling sequential user browsing increases consistency with human preferences, (ii) modeling novelty and variety increases consistency with human preferences, and (iii) modeling image visual saliency increases consistency with human preferences. We observe this behavior on three different datasets including MS-COCO captions [21], the localized narratives [70] and a set of aggregated prompts generated by humans. We release our code at https://github.com/Narabzad/Set-Based-Text-to-ImageGeneration.

## 2 Related work

**Text to Image Generation Evaluation** Despite the recent advances in text guided image generation models [31, 33, 49–51, 71–73, 77, 78, 91, 98], evaluating such systems have not been studied extensively. The lack of evaluation tools for such systems has been identified as a serious issue within the community [24, 55, 62, 63, 67]. The Inception Score (IS) measures properties of *unconditional* generated images for (i) low-entropy across image classes, and (ii) high entropy across image classes over a population of generated images [79]. However, the Inception Score (IS) is susceptible to gaming a perfect score by creating one image for each of the base embedding classes. Therefore, while better performance is possible, the metric does not fully capture the potential for improvement. Additionally, IS is appropriate for assessing only certain properties of a general, unconditional image generation model and is not suitable for TTI evaluation. As a mathematical refinement of IS, the Fréchet Inception Distance (FID) compares the distribution of Inception embeddings of real and generated examples [26, 37]. A lower FID indicates a higher similarity between the real and generated distributions, and thus more realistic images. However, like IS, FID was designed for unconditional image generation, making its use for TTI evaluation somewhat contrived [4]. Although other methods, like Kernel Inception Distance [9], have proposed further mathematical improvements, FID remains the dominant evaluation metric

[62]. Multiple studies have demonstrated that population-level metrics like FID are not correlated with human preferences [54, 62, 66]. In particular, Otani et al. [62] rigorously analyzed TTI evaluation, finding weak alignment with human behavior and perception as a primary cause of this inconsistency.

A second group of metrics evaluates the alignment between generated images and the user prompt. Some metrics achieve this by algorithmically captioning the generated image and then computing text similarity to the user prompt [23, 92]. Other metrics measure semantic object accuracy, which assesses how well objects in the generated image match terms in the caption [38, 39, 97]. This group of evaluation metrics is limited by the performance of a second model (e.g., caption generation, object detection model, or visual question answering models) and usually does not show a high correlation with human-generated prompts, as reported in previous works [23, 65, 78, 98]. In addition, they lack the flexibility to be applied to *set-based* evaluation of TTI systems, which is the focus of our work. Even state-of-the-art caption generation models struggle to produce distinct captions for images generated from the same TTI system for a given prompt [34, 43, 85]. Our experiments demonstrate that even with a SOTA pretrained image captioner model such as VL-T5[22], there are significant overlaps between the generated captions for images generated from an individual TTI system for a single prompts. Consequently, the evaluation metric would yield the same relevance score for all generated images within the set. As a result, these metrics are not effective in evaluating TTI systems on set-based instances, which is the focus of our work. In a recent study [95], the authors introduced a method to score and evaluate human preferences. Similarly, in [46] the authors collected a preference-based dataset to train a CLIP-based scoring function so called as PickScore to predict human preferences on individual pair of generated images for a given prompt. However, their focus was on individual generated images rather than considering set-level evaluation, which crucially incorporates diversity alongside relevance. In our work, we only focus on TTI evaluation metrics that are applicable to set-based TTI generation evaluation.

There have been also attempts to develop evaluation pipelines for TTI models [7, 39, 40, 46]. For instance, Petsiuk et al. [68] created a set of fifty challenging tasks and applications for state-of-the-art TTI models. As a result, several curated benchmarks for TTI evaluation have been developed. Moreover Yu et al. [98] introduced PartiPrompts, a comprehensive benchmark of 1,600 English prompts covering 12 different categories and 11 challenge aspects. In PartiPrompts, similar to Xu et al. [95] human annotators were asked to select the better generated image from two models (one image per model). Although the majority of preference-based TTI evaluation focuses on pairs of single images, Saharia et al. [78] conducted preference-based evaluations between sets of images generated by two models. To the best of our knowledge, their collected dataset is the only work that conducted evaluation on set of images. In our work, we follow the same concept and aim to conduct experiments comparing sets of generated images, albeit on a larger scale.

**Diversity** Measuring the diversity of image sets has a long history in multimedia retrieval [1, 56, 87]. For example, the MediaEval 2017 benchmark introduced image search result diversification in the context of photo retrieval [99], arguing that diversity within

image search results is crucial because it (i) addresses the needs of different users, (ii) tackles requests with unclear information needs, (iii) widens the pool of possible results for increasing the performance, and (iv) reduces redundancy in the results. We adopt similar desiderata motivated from the the precise ideation context in which TTI is often used. While MediaEval could use relevance and diversity unmodified IR metrics because images were explicitly rated, TTI needs to impute these ratings because images are generated.

## 3 Method

### 3.1 Problem Definition

We define the *TTI Task* as: Given a prompt or query $q$, a TTI system $s$, generates an $w \times h$ matrix $\mathcal{X}$ where each element is an image. If the TTI interface returns a ranked list of length $k$, we generate a $1 \times k$ matrix and similarly when a grid is required, the TTI system presents the generated images in a grid view of $m \times n$ images. Further, we define the *TTI Evaluation Task* as: An evaluation metric $\mu$ is a function that, given an arrangement of generated images $\mathcal{X}$, a prompt $q$, and side information $u_q$ about the image utility (e.g. an example target image), computes a scalar value where a higher value indicates better performance of system. We summarize the performance of system $s$ over a space of prompts $Q$ with the expected value of $\mu$ over the distribution of queries in $Q$, $\mathbb{E}_{q \sim Q}[\mu(s(q), u_q)]$.

### 3.2 Designing a Metric for TTI

As mentioned in Section 1, in the context of creative design, users leverage TTI systems to generate design candidates to support ideation. Dominant theories for measuring the quality of an ideation process focus on four criteria [35, 45, 80]. *Fluency* refers to the total number of relevant items generated. *Variety* refers to the number of unique types of relevant items generated. *Novelty* refers to how different relevant items are from all previously generated items. Finally, *quality* refers to the degree of relevance of generated items. We are interested in designing metrics that capture these different dimensions of ideation effectiveness. Here, we will review relevant concepts from information retrieval evaluation as they pertain to TTI evaluation.

**A General Model of User Behavior:** Let $\mathcal{X}$ be the matrix of $k = wh$ images generated for a specific prompt, where we index each generated image from 1 to $k$. When inspecting a grid, we define the trajectory $\pi$ as be the specific sequence of inspected images represented as a permutation of $[1, k]$, where $\pi_i$ refers to the $i$th examined image index. We begin our metric development in the context of classic IR evaluation where users inspect a one-dimensional ranked list in a serial, deterministic order. We will return to grid-based interfaces at the end of this section.

We also have available information about the *relevance* of an image to the user for a prompt $q$. Let $f^*(x) \in [0, 1]$ represent the relevance of an image $x$ to the user.

Given a trajectory $\pi$, we define a family of metrics based on how the user might engage with the images in linear order. The *position-based model* models the probability of a user inspecting the image at rank position $i$ in the trajectory as $\gamma^{i-1}$, where $\gamma \in [0, 1]$ is a free parameter controlling the depth the user is likely to reach [27, 75]. The *cascade model* models a user who, after inspecting

a relevant image, might be satisfied and terminate their scan [18, 19, 27]. Let the probability that a user is satisfied by an image be a monotonically increasing function of its relevance, $g^*(f^*(x))$, which we represent as shorthand using $g^*(x)$ for clarity. In this case, the probability that they reach rank position $i$ is $\gamma^{j-1} \prod_{j=1}^{i-1}(1 - g^*(\pi_j))$. In the remainder of this section, we will demonstrate how to use these models to design metrics capturing fluency, quality, novelty, and variety.

**Fluency:** To measure fluency, we define a metric based on the expected number of relevant images the user will see. For the position-based model, rank-biased precision or RBP [57], is defined as,

$$\text{RBP}(\pi) = \sum_{i=1}^{k} f^*(\pi_i) \gamma^{i-1} \qquad (1)$$

Under the cascade model, extended expected reciprocal rank (ERR) metric [18], is defined as,

$$\text{ERR}(\pi) = \sum_{i=1}^{k} f^*(\pi_i) \gamma^{i-1} \prod_{j=1}^{i-1}(1 - g^*(\pi_j)) \qquad (2)$$

**Quality:** Unlike classic IR evaluation, we cannot *a priori* judge the relevance of all possible images, we leverage a wide variety of side information to estimate the relevance of generated images. In general, this could include information about user preferences, image attributes, or anything else helpful to estimate relevance. In our study, we assume that we have access to one or more example images known to be relevant to the prompt. Given a generated image $x$ and a relevant image $\hat{x}$, we estimate the relevance of $x$ as the following where $\phi$ is an image embedding function, for example based on the activation of an interior layer of a neural network:

$$f^*(x) = \langle \phi(\hat{x}), \phi(x) \rangle \qquad (3)$$

Carterette and Allan [15] uses a similar approach to label unlabeled documents in classic IR evaluation.

**Variety and Novelty:** In TTI evaluation, for a given prompt, we only have a few example images and therefore do not have information to measure novelty of a generated image to the user or in general. However, we can measure the novelty of an image to the user *while they scan the list*. Given the similarity of novelty within a list to variety within a list, we consider both of these criteria together. In situations where we have additional information about the novelty of an image to a user or in general, we suggest these criteria be decoupled. In order to measure novelty within a list, we leverage existing methods designed to measure diversity. Similar to Maximum Marginal Relevance [36], we model novelty by discounting the relevance of an image based on previously seen images. The discount factor for the $i$th image in $\pi$ is defined as the following where $\eta(i, \pi)$ can be used to discount $f^*(\pi_i)$ which we can then use in either RBP or ERR.

$$\eta(i, \pi) = 1 - \max_{j \in [1, i-1]} \langle \phi(\pi_i), \phi(\pi_j) \rangle \qquad (4)$$

**Expected Metrics over Trajectories:** In more general grid-based interfaces, we cannot assume that users will all follow a deterministic trajectory when presented with an arrangement of images [30]. Users may inspect images in arbitrary trajectories based on their position and attractiveness. As such, we can compute the expected metric value over all possible trajectories. Let
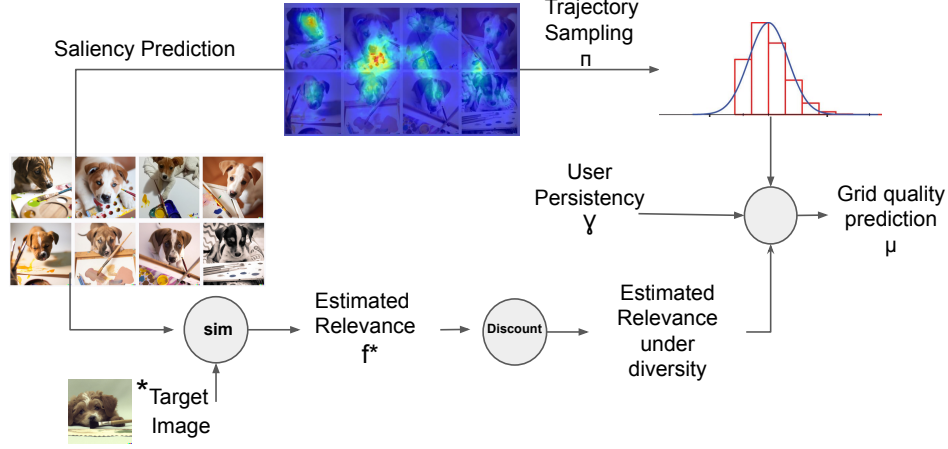
**Figure 1: Overview of our proposed approach**

**Table 1: Examples of prompts from `MS-COCO`, `LN-COCO` and `Prompts` dataset as well as average and standard deviation $\sigma$ of length of prompts in each dataset.**

| Dataset | # | #Words | $\sigma$ | Example Prompt |
|---|---|---|---|---|
| MS-COCO | 500 | 10.3 | 2.2 | A herd of cows standing on a grass covered hillside. |
| LN-COCO | 500 | 42.8 | 19.2 | In this picture we can see three cows standing on the grass. There is a tree and few mountains are visible in the background. |
| Prompts | 500 | 11.6 | 9.4 | Origami cow flying over the moon. |

$\Pr(\pi)$ be the probability that the user scans the images in the order represented by $\pi$; in other words, $\Pr(\pi)$ is a multinomial distribution over all permutations in $S_k$. Although a naïve model might consider a uniform distribution over trajectories, we know that users tend to be more attracted to certain images based on their position and visual features [30]. Inspired by the observation that users tend to look at the most salient images first [59, 83], we model $\Pr(\pi)$ using a Plackett-Luce model based on predicted image salience [69]. Let $\sigma(x)$ be the visual salience of $x$ represented as a positive scalar value. After normalizing this value over all images to sum to one, we iteratively sample an image from this multinomial *without replacement*, generating a sequence of images (i.e. $\pi$). This process can be executed efficiently by using the Gumbel-softmax method [12]. Figure 1 displays a general overview of our proposed approach for evaluating TTI systems.

## 4 Experimental Setup

### 4.1 TTI Systems

To run experiments and validate how our proposed evaluation metrics can distinguish between the two sets of generated images, we took three state-of-the-art TTI systems [78, 98]. In order to preserve the anonymity of the systems and emphasize evaluation, in this paper, we refer to these as systems A, B and B'. B' is the smaller version of system B, with fewer number of parameters. In prior work, systems A and B have shown competitive performance and consequently, the smaller version of system B, system B', has shown relatively worse performance compared to system A. The assumption that a model with fewer parameters has worse performance compared to a larger model of the same class has previously been used for evaluation purposes [67].

### 4.2 Datasets

To conduct our analysis, we require sets of prompts with associated prior information. As discussed in Section 3.2, each prompt is linked to at least one relevant target image. We utilized the COCO dataset [53], which provides two distinct representations of information needs: 1) COCO captions (`MS-COCO`) [21] and 2) Localized Narratives (`LN-COCO`) [70]. Additionally, we collected a third dataset, `Prompts`, comprising of prompts entered by users of a TTI demo.

The examples in Table 1 highlight the difference between prompts used for these experiments for the same image. As shown in the first two rows of this Table, the same image was described with 10 words when using `MS-COCO` captions and 24 words when using `LN-COCO` as prompts. We demonstrate how a same image was described with 10 words when using `MS-COCO` captions and 24 words when using `LN-COCO` as prompts. We are interested in investigating whether our proposed metrics and baselines are robust with respect to the lengths of the prompts for the same prior information. While the average number of words in `Prompts` and `MS-COCO` seem close (11.6 vs. 10.3), the standard deviation of the number of words in the prompts of these two datasets differs significantly. The standard deviation of the number of terms in `Prompts` is 9.41, while this number for `MS-COCO` is only 2.21. From Table 1, we can see that prompts in the `MS-COCO` dataset are consistently short, while in the `Prompts` dataset, the length of the prompts varies more. We conclude that prompts in the `MS-COCO` dataset are consistently short, while in the `Prompts` dataset, the length of the prompts varies more.

**COCO Captions** The COCO dataset is widely used to evaluate various deep learning and computer vision tasks, including TTI systems [20, 31, 53, 78, 81, 98, 101]. The MS-COCO captions dataset contains over 200K images, each annotated with five captions. For this paper, we randomly sampled 500 images from the MS COCO
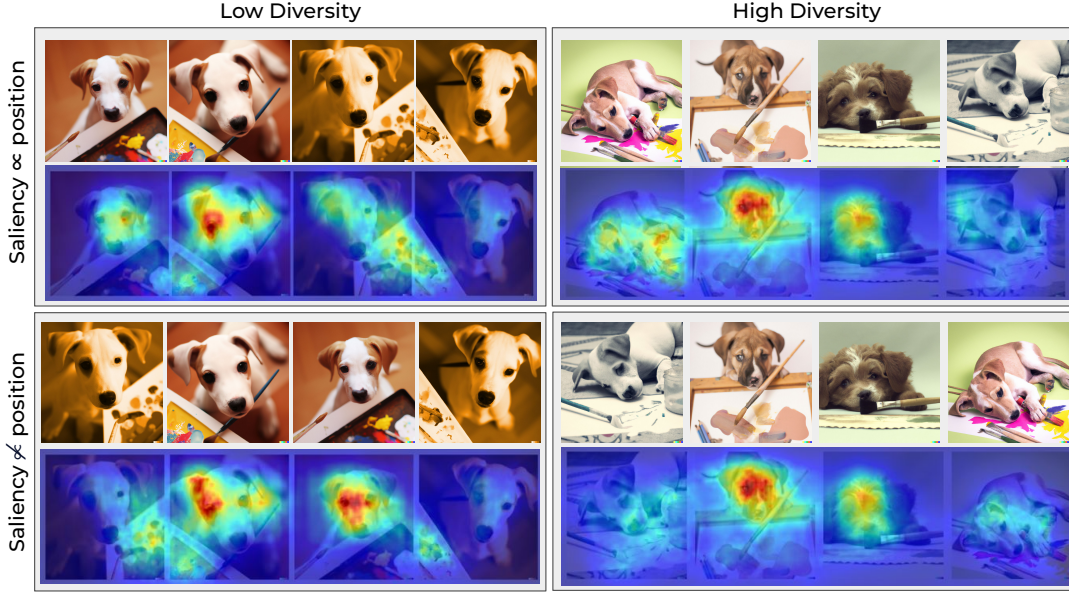
**Figure 2: Four image grids and associated salience maps for the prompt 'a cute puppy is painting'. Images grids in the left column have lower diversity than those in the right column as reflected by the different positions and breeds on puppies. In image grids in the top row, images tend decrease in salience from left to right; in the bottom row, there is less of a relationship between position and salience.**

2017 validation set. For each image, we randomly selected one of the captions as the input prompt for experiments and used the associated image as the target image.

**Localized Narratives** Recent work has raised concerns that COCO captions might not be realistic surrogates for human prompts [98]. Therefore, we use Localized Narratives dataset which are a subset of the MS-COCO dataset, and their associated text is four times longer than COCO captions on average [70]. Inspired by previous work [49, 100, 101], we also conducted experiments with generated images from more detailed descriptions of images (localized narrations) and examined how the longer version of prompts affects the performance of TTI systems.

**Prompts** Describing an image through captions or detailed narrations is different from issuing a prompt to address an information need in various contexts. To ensure that our evaluation metric is representative of prompts that appear in practice, we also analyzed performance on human-generated prompts. We collected a third dataset including 1,500 prompts from real users, referred to as `Prompts`, consisting of prompts entered by users of a TTI demo. This dataset is more representative of the types of prompts that TTI systems will encounter in practice, and they cover a wide range of topics and styles. Some examples of prompts in `Prompts` are *"A futuristic city with lots of greenery", "A small cabin in the woods"*, and *"A cozy reading nook in a library"*. The TTI demo was used for generating images for various purposes, including research, design, and presentations. We also recorded the generated grids from two TTI systems, `A` and `B`, and collected feedback on individual images, including positive and negative signals, such as thumbs up and down, as well as implicit signals, such as downloads, copies, or shares. We used both implicit and explicit positive feedback to create a set of target images for each prompt, selecting only those prompts that were tried on both TTI systems and received positive feedback on both. This resulted in a set of target images, consisting of at least one image from each of the two TTI systems, for each prompt. We randomly sampled 500 such prompts that satisfied these requirements.

## 4.3 Metric Variations

We present our six different metric variations in Table 2. RBP and ERR metrics are defined in Equation 1 and 2 respectively. The subscripts $\eta$ discounts the relevance function as explained in Equation 4. When $\pi$ is sampled based on uniform distribution the metric has a superscript $u$ otherwise the $\pi$ is sampled using the normalized image saliency and unless specified, we adopt a value of $\gamma$ equal to 0.9, as suggested in the literature [14]. We compute the average of the metric over 100 sample trajectories for each prompt. The similarity between generated images and target images is calculated using the cosine similarity of the embedded vectors of images obtained from Inception V3 which has been widely used in downstream vision tasks and evaluation methodologies [8, 52, 60, 72, 86, 90, 94].

## 4.4 Baseline TTI Evaluation Metrics

In addition to FID, we develop a simple diversity-based metric, *Diversity*, to assess the set-based diversity of TTI output. *Diversity* measures the average pairwise similarity between all generated images. A higher pairwise similarity rate within a grid indicates *lower* diversity. We compute pairwise similarity using cosine similarity of the embedded vectors of images obtained from Inception V3.

**Table 2: Different variations of our proposed metrics**

| Metric | User Model | Relevance | $Pr(\pi)$ | Metric | User Model | Relevance | $Pr(\pi)$ |
|---|---|---|---|---|---|---|---|
| RBP | Position-based | $f^*(x)$ | Saliency | ERR | Cascade-based | $f^*(x)$ | Saliency |
| $RBP_\eta^u$ | Position-based | $f^*(x) \times \eta(i, \pi)$ | Uniform | $ERR^u_\eta$ | Cascade-based | $f^*(x) \times \eta(i, \pi)$ | Uniform |
| $RBP_\eta$ | Position-based | $f^*(x) \times \eta(i, \pi)$ | Saliency | $ERR_\eta$ | Cascade-based | $f^*(x) \times \eta(i, \pi)$ | Saliency |

## 4.5 Modeling Salience

Although there are no existing saliency models or gaze data available for a grid of images, we can use models trained on webpages, which usually contain multiple images (or even a grid of images sometimes). There are several existing free viewing gaze data on webpages [16, 82], and among them, [16] is the latest and largest, with gaze data on 450 webpage screenshots, collected from 41 people with eye trackers. 450 webpages are a relatively small data set, and insufficient to train the webpage saliency model. We follow the typical training paradigm in saliency modeling area: first train the model with Salicon data [42], a large scale saliency data set on 10K natural images, then fine-tune it with webpage gaze data. The loss functions we use is a combination of KLD and NSS [13], both of which are popular loss function for saliency models. Our webpage saliency architecture essentially follows the SimpleNet idea in [74] with some modification and simplification. More specifically it takes a MobileNet V3 (pretrained on ImageNet classification data) as backbone, and extracts embeddings from 4 layers of conv 2, 4, 6, 8. On each embedding, we apply two conv layers, where the first layer has a kernel $3 \times 3$, number of channels matching the input embedding, max pooling 3 and relu operator and the second layer has a kernel $1 \times 1$, 1 channel and relu operator. Then the output of each branch is converted back to the resolution of the input with a bilinear resize operator. The outputs of the 4 branches are then summed to 1 channel, followed up by a sigmoid function to generate the output saliency map. Finally, we average the predicted saliency of individual images within the grid and use it as initial distribution to sample $\pi$ [1].

## 4.6 Data Annotation

*4.6.1 Annotation Process.* We curated a dataset of human preferences over sets of generated images for prompts from the three above-mentioned datasets. An annotation study was conducted using an internal platform, where annotators selected their preferred set of images for each prompt. Based on prior work in task-based evaluation of information retrieval systems [11], we grounded the annotation in a simulated work task provided to annotators as,

*" You and your coworkers are trying to come up with an image for a project presentation. Together, you all have come up with a description of the image, which we will refer to as a 'prompt' (shown in blue below). Two designers, X and Y, have sketched possible images for the group to decide on the images to use. Which designer's sketches would you prefer to present to the group to decide on the presentation image?"*

The instructions, along with a prompt, were displayed at the top of the assessment interface. Two sets of eight images were

presented in a $2 \times 4$ grid view, one from TTI system X on the left and one from another TTI system Y on the right, immediately below the guidelines and the prompt. Annotators rated which of the grids they preferred on a five-point Likert scale: 1) X is much better than Y, 2) X is somewhat better than Y, 3) X and Y are the same, 4) Y is somewhat better than X, and 5) Y is much better than X.

To ensure the quality of the collected annotations, each pair of sets of generated images was annotated three times with different annotators. We used the consensus agreement among all three judgments to generate the final annotation results. We ran preference-based annotation between each pair of systems, i.e., A vs. B, A vs. B', and B vs. B', on 500 prompts from each of the MS-COCO, LN-COCO, and Prompts datasets.

In total, we collected 13,500 annotations for our dataset. Additionally, we conducted 100 quality check tests in which we presented a generated set of images for a given prompt and a random set of generated images from a different prompt. We assumed that the randomly generated set of images would always be less relevant. We are pleased to report that all of the quality tests were passed with the consensus results. Furthermore, the crowd sourcing platform we utilized had their own internal quality checks and qualification tests which were passed at the most satisfactory level. On average, each annotation took 106 seconds to complete and although we paid the raters hourly, on average we paid 0.50 USD per annotation.

*4.6.2 Agreement Rate.* Three different annotators judged each pair of sets of generated images. We computed Fleiss' $\kappa$ independently for each dataset. In Table 3, we provide the agreement rate between the annotators when annotating different pairs of systems on the three datasets separately, as well as the total agreement on all the annotated pairs of grids. We also report the agreement rate on a three-point scale where we collapse ratings in the same *direction* (e.g., considering 'A is much better than B' and 'A is somewhat better than B' as one option 'A is better than B'). As shown in this Table, the agreement rate between the three annotators was quite promising, and the majority of the prompts received at least two out of the three annotator's agreement. Even on a 5-scale consensus result, which is harder to achieve higher agreement on, only 13.6% of the prompts from MS-COCO, 12.3% of prompts from LN-COCO, and 12.6% of prompts from Prompts dataset did not have any consensus agreement between the three annotators. We qualitatively observed some of the prompts that received 0 agreement rate between the three annotators and noted that in these special cases, both of the models generated a relevant set of images, and thus the annotator's selection became less robust.

When looking at agreement across annotators for grades on our Likert scale, we observe moderate agreement MS-COCO ($\kappa = 0.57$) and LN-COCO ($\kappa = 0.58$) and fair agreement for Prompts ($\kappa = 0.33$). This suggests that systems are easier to precisely distinguish in

---

[1]We release our code at https://github.com/Narabzad/Set-Based-Text-to-ImageGeneration
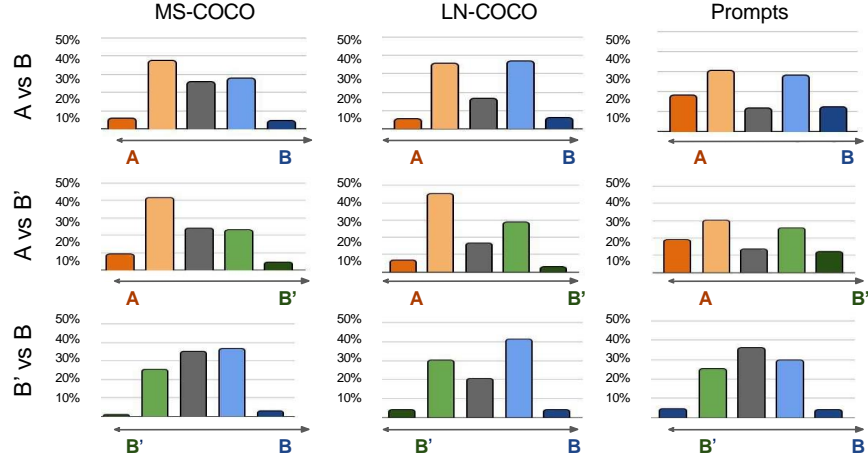
**Figure 3: Annotation results of TTI systems A (orange), B (blue) and B' (green) on the three datasets. Given the TTI systems X/Y are on the left/right side of the arrows beneath a sub-figure, the bars in the sub-figure present if "X is much better than Y", "X is somewhat better than Y", "X and Y are the same", "Y is better than X" and "Y is much better than X" consecutively.**

MS-COCO and LN-COCO while the Prompts dataset is more challenging for annotators. We also computed agreement in *direction*, collapsing the bottom two and top two grades. When computing the directional agreement, we observe substantial agreement for LN-COCO ($\kappa = 0.62$) and moderate agreement for MS-COCO ($\kappa = 0.56$) and Prompts ($\kappa = 0.41$). We also observed that, despite perhaps being more challenging when evaluating these specific systems, the Prompts dataset has equal or better majority agreement compared to MS-COCO and LN-COCO. From an evaluation perspective, the difficulty in distinguishing systems could arise from, for example, harder to distinguish pairs of image sets or more ambiguous prompts. If we had observed substantially lower agreement, we would question whether these systems were in fact different in terms of the prompts in the Prompts dataset.

Additionally, by comparing the agreement rate on the three datasets, we observe the agreement rate on the Prompts dataset is significantly lower than the other two datasets. For instance, while 57.8% of the prompts in the MS-COCO dataset and 61.8% of prompts in the LN-COCO dataset had 3 out of 3 consensus agreement on a 5-scale annotation result, only 28.5% of the prompts in the Prompts dataset had full agreement rate between the 3 annotators. Obviously, the agreement rate increases when we only consider the direction of the options and neglect the magnitude. For instance, the 3/3 agreement rate increases from 57.8% to 63.5% for MS-COCO, 61.8% to 68.6% for LN-COCO, and from 28.5% to 45.8% in the Prompts dataset. However, we still see a notable difference between the three datasets in the agreement rate. We hypothesize that this difference could be because the prompts in the Prompts dataset are more challenging and point to more specific concepts and design patterns, such as the vibes, the quality, style, etc. However, the prompts in the LN-COCO and MS-COCO datasets do not elaborate on such specifications, and thus they obtained a higher agreement rate between the three annotators.

## 5 Results and Validation

### 5.1 Annotation Results

We present the results of our annotation experiments on the five-point Likert scale in Figure 3 for the three datasets. These results demonstrate that the preference rate between each pair of systems is highly dependent on the dataset (i.e., the composition of input prompts). We note that the MS-COCO prompts are the shortest and simplest ones for TTI systems to address. In contrast, the Prompts dataset and the LN-COCO prompts are longer and more challenging, as shown in previous work [78, 98]. Observing the first row of Figure 3 (A vs B), we note that when prompts are simpler (i.e., MS-COCO), both systems A and B perform well. In fact, in 26% of the comparisons, annotators rated both systems as "A and B are the same." However, for more challenging datasets, we observed ties amongst 17% (LN-COCO) and 12% (Prompts) of the comparisons. Consequently, the percentage of extreme choices (i.e., when one system is preferred "much" better than the other) is lowest in MS-COCO, which confirms our previous observation that both systems perform similarly, making it difficult for annotators to distinguish between them. On the other hand, in the Prompts dataset, annotators tended to select extreme choices (i.e., one model is "much better" than the other one) more frequently than the other two datasets. Therefore, when the prompts become more challenging, it is easier for annotators to discriminate between two models, as it is more likely that one of the systems fails to satisfy the complex prompt. When the prompts are simple, we hypothesize that both models perform well, and annotators do not perceive many differences between the two sets of generated images.

When comparing the first row with the second row in Figure 3 and keeping in mind that B' is the version of B with fewer parameters, we observe that system A was preferred more frequently in the A vs B' comparison rather than in the A vs B comparison across all three datasets. For example, in the MS-COCO dataset, A was rated as much or somewhat better than B in 43% of the prompts, while A was preferred over B' in more than 50% of the prompts.

**Table 3: Agreement rate between the three annotators on each of the datasets and between each pairs of TTI systems on a 5-level and 3-level Likert scale. Fleiss' $\kappa$ is reported on both 3-levels and 5-levels consensus.**

| | Dataset | 5-levels Consensus | | | | 3-levels Consensus | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0/3 | 2/3 | 3/3 | $\kappa$ | 0/3 | 2/3 | 3/3 | $\kappa$ |
| A | MS-COCO | 15.6% | 24.5% | 59.9% | - | 13.7% | 21.7% | 64.6% | - |
| vs | LN-COCO | 10.6% | 26.9% | 62.6% | - | 9.3% | 18.9% | 71.8% | - |
| B | Prompts | 11.8% | 60.7% | 27.5% | - | 4.0% | 44.3% | 51.7% | - |
| A | MS-COCO | 12.6% | 25.7% | 61.7% | - | 10.8% | 23.8% | 65.4% | - |
| vs | LN-COCO | 12.7% | 25.4% | 61.8% | - | 10.1% | 21.5% | 68.4% | - |
| B' | Prompts | 12.2% | 59.5% | 28.3% | - | 4.6% | 43.7% | 51.7% | - |
| B | MS-COCO | 12.6% | 25.7% | 61.7% | - | 18.2% | 21.4% | 60.4% | - |
| vs | LN-COCO | 13.7% | 25.2% | 61.1% | - | 12.4% | 22.1% | 65.5% | - |
| B' | Prompts | 13.6% | 56.7% | 29.7% | - | 8.0% | 58.1% | 33.9% | - |
| | MS-COCO | 13.6% | 25.3% | 61.1% | 0.57 | 14.2% | 22.3% | 63.5% | 0.56 |
| All | LN-COCO | 12.3% | 25.8% | 61.8% | 0.58 | 10.6% | 20.9% | 68.6% | 0.62 |
| | Prompts | 12.6% | 59.0% | 28.5% | 0.33 | 5.5% | 48.7% | 45.8% | 0.41 |
| Average Consensus | | 12.8% | 36.7% | 45.1% | - | 10.1% | 30.6% | 59.3% | - |

This increase in the preference rate of A from the comparison to B versus the comparison to B', indicates that, as expected, the larger model B was more competitive with A than the smaller model B'. It is also interesting that the option "X and Y are the same" received the highest rate when comparing TTI system B with B'. Especially for the Prompts dataset, the two systems are almost indistinguishable and were labeled as the same in over 36% of the prompts. In comparison, between systems B vs B', we also note the smallest difference in preference between the two sides, with less than a 4% difference between when B was preferred over B' and when B' was preferred over B.

## 5.2 Agreement with Human preferences

Given two sets of images generated by systems $s$ and $s'$, we expect $\mu(s(q), u_q) > \mu(s'(q), u_q)$ if human annotators preferred the set of generated images from $s$ over $s'$. To ensure the reliability of agreement between annotators, we only report the agreement rates of the evaluation metrics on prompts where all three annotators fully agreed on the annotations. The agreement rates in Table 4 indicate the percentage of the prompts where annotators rate system X as better than system Y *and* our metric also computes $\mu(X(q), u_q) > \mu(Y(q), u_q)$. Since FID is not capable of quantifying the quality of a single set of generated images for a given prompt, it is impossible to report the agreement rate between annotators on a per-sample basis. We also report the results of Wilcoxon paired statistical significance tests with p-values < 0.05 between the annotations and the measured performance by $\mu$. As shown in Table 4, all six of our proposed metrics are able to show statistically significant agreement rates with annotations when comparing A vs B' on all three datasets. However, they do not show statistically significant agreement with annotators when comparing B vs B'. We hypothesize that since B' is a version of B only varying by the number of parameters, their performance may indeed be quite similar. Therefore, the non-significant agreement rate in this comparison could be the result of actual comparable performance between the two systems. The third row of Figure 3 confirms this since the two systems have the highest "B *and* B' *are the same*" rate (the gray bars) across all three datasets.

We additionally note that on the two most challenging datasets (LN-COCO and Prompts), we see statistically significant agreement rates with annotators when comparing TTI model A vs B. We note that (i) MS-COCO prompts are the most simple prompts among the three datasets, and (ii) the MS-COCO dataset have been widely used for training TTI systems, and so, TTI models are more familiar with the structure of simple prompts and images. However, for more complex prompts, it is more likely that one of the models fails to generate a relevant set of images, making it easier to assess their performance. As shown in Table 4, our proposed metric $ERR_\eta$ outperforms the other evaluation metrics on the Prompts dataset. We achieved a 59.3% agreement rate with annotators when comparing A vs B and a 63.1% agreement rate when comparing A vs B'. This metric also shows favorable agreement rates on the other two datasets. Additionally, the position-based metric $RBP_\eta$, which considers variety and novelty, shows the highest agreement on the MS-COCO dataset. In general, novelty-based metrics were able to outperform other metrics on all three datasets, confirming the impact of considering variety and novelty in TTI evaluation metrics.

## 5.3 Ablation Study

We studied the impact of $Pr(\pi)$ (Section 3.2) on the agreement rates by comparing $RBP_\eta$ with $RBP_\eta^u$ and similarly comparing $ERR_\eta$ with $ERR_\eta^u$. We observe in Table 4 that $ERR_\eta$ shows a superior agreement rate compared to $ERR_\eta^u$ on the majority of datasets when the agreement rate is statistically significant. Additionally, $RBP_\eta$ also shows a greater agreement rate compared to $RBP_\eta^u$ on two out of the three datasets. By comparing the impact of $Pr(\pi)$ across the three datasets, we hypothesize that sampling $\pi$ from the saliency distribution of the grid works best as the prompt sets become more challenging and it has the most positive effect on the Prompts dataset and less positive effect on the easier prompt sets. We have also included the agreement rate of the *Diversity* metric. However, it failed to show any statistically significant agreement with the human annotators. Nonetheless, by comparing the performance of metrics with and without the variety and novelty components, we can conclude that incorporating variety and novelty in TTI evaluation

**Table 4: Agreement rate of the metrics with human annotations. Statistically significant agreement with Wilcoxon paired test and p-value < 0.05 are shown with * symbol. Bold: highest statistically significant agreement in the column.**

| | MSCOCO | | | LNCOCO | | | prompts | | |
|---|---|---|---|---|---|---|---|---|---|
| | A vs B | A vs B' | B vs B' | A vs B | A vs B' | B vs B' | A vs B | A vs B' | B vs B' |
| Diversity | 47.2% | 41.9% | 53.1% | 45.3% | 42.1% | 45.7% | 45.9% | 46.3% | 47.1% |
| rbp | 53.7% | 55.3%* | 55.5% | 53.1%* | 64.1%* | 52.1% | 56.9%* | 60.2%* | 54.2% |
| novrbp | **53.8%** | **61.2%*** | 51.7% | 52.5%* | 58.8%* | 54.2% | 58.5%* | 60.7%* | 54.2% |
| urbp | 53.3% | 58.9%* | 52.9% | 54.2%* | 60.0%* | 53.4% | 57.7%* | 60.2%* | 49.5% |
| err | 52.2% | 55.4%* | 51.7% | 50.0%* | 60.1%* | 54.2% | 57.7%* | 61.5%* | 53.5% |
| noverr | 52.2% | 60.4%* | 50.9% | 53.1%* | 60.8%* | 56.2% | **59.3%*** | **63.1%*** | 54.2% |
| uerr | 52.2% | 58.4%* | 52.9% | **55.2%*** | **62.6%*** | 52.7% | 58.1%* | 61.5%* | 52.3% |

**Table 5: Distinguishing images from the Preferred ($P$) and Not-preferred ($N$) set of grids in the preference-based comparisons. The lower FID in each comparison indicates the better set. For ERR and RBP, higher value indicates better predicted performance. For the sake of easier comparisons, we multiply the output of our metrics by 100. Bold: The better predicted.**

| Dataset | Metric | A vs B | | A vs B' | | B vs B' | |
|---|---|---|---|---|---|---|---|
| | | Preferred (P) | Not-preferred (N) | Preferred (P) | Not-preferred (N) | Preferred (P) | Not-preferred (N) |
| MS-COCO | FID | **27802.9** | 28991.4 | **23647** | 26586.6 | **26673.1** | 28757.2 |
| | ERR | **98.75** | 98.69 | **98.79** | 98.75 | **98.72** | 98.71 |
| | ERR$_\eta$ | **91.23** | 90.99 | **91.54** | 91.35 | **91.35** | 91.25 |
| | RBP | **89.01** | 88.60 | **89.36** | 88.96 | **88.77** | 88.69 |
| | RBP$_\eta$ | **80.73** | 80.40 | **81.04** | 80.75 | **80.64** | 80.54 |
| LN-COCO | FID | 24191.6 | **21983.9** | 22671.9 | **21237.3** | **23647.0** | 26586.6 |
| | ERR | **96.56** | 96.34 | **96.85** | 96.38 | **96.23** | 96.08 |
| | ERR$_\eta$ | **90.8** | 90.27 | **91.43** | 90.35 | **90.08** | 89.76 |
| | RBP | **89.58** | 89.05 | **90.39** | 89.04 | **88.66** | 88.35 |
| | RBP$_\eta$ | **64.65** | 64.27 | **65.04** | 64.29 | **64.2** | 63.96 |
| Prompts | FID | **3451.5** | 4748.6 | **3372.1** | 4792.3 | **9012.0** | 9133.8 |
| | ERR | **96.55** | 96.33 | **96.48** | 96.29 | **96.43** | 96.38 |
| | ERR$_\eta$ | **90.91** | 90.41 | **90.76** | 90.27 | **90.55** | 90.49 |
| | RBP | **89.47** | 88.86 | **89.29** | 88.76 | **89.1** | 88.99 |
| | RBP$_\eta$ | **64.85** | 64.5 | **64.73** | 64.39 | **64.63** | 64.61 |

metrics improves the agreement rates of both the position-based and cascade-based metrics with the human annotations [5].

## 5.4 Comparison with FID

Although we cannot measure prompt-level agreement for FID, we can measure the population-level agreement. If comparing two systems $X$ and $Y$, we can use the set of all images in the Preferred grids *regardless of the source system*; call this set ($P$). Similarly, we have the set of images in Non-preferred grids ($N$). We measure FID between the following sets: 1) set of target images and all the generated images from the Preferred grids ($P$) and 2) set of target images and all the generated images in Non-preferred grids ($N$). We expect that the set of Non-preferred grids to show higher FID (lower similarity) with the target images. Our results (Table 5) show that, while FID is able to measure better performance (lower distance with target images) on the simplest set of prompts (i.e., MS-COCO), it fails to detect the better-performed set for challenging prompt sets. In fact, for LN-COCO, FID indicated that the Non-preferred sets of generated images had performed better when comparing both A vs B and A vs B'. However, it correctly measured a lower FID score for the preferred set in the comparison of B and B'.

We can compare the population-level consistency of FID with that of our proposed metrics. Similar to the FID analysis, we report the average of our metrics on both P and N sets of generated images. Unlike FID, all of our metrics were able to assign higher values when comparing any pairs of the TTI systems and on all the three datasets. All four of our metrics consider the generated images from

Preferred grids (P) as the better set compared to the images from Non-preferred grids (N).

## 6 Conclusion

Because reliable offline evaluation metrics for TTI systems is critical for benchmarking purposes in the community [62], we proposed a set of evaluation metrics for set-based TTI inspired by grounding ideation criteria in traditional IR evaluation methods. We use existing and novel approaches to capture fluency, variety, novelty, and quality, while incorporating visual salience as a fundamental feature of image layouts. Moreover, to the best of our knowledge, this is the first work to explore assessing generated images as a *set*. We demonstrated the effectiveness of our family of metrics against the popular FID metric, showing that ideation-based measures were better aligned with human preferences. Although our results are positive, we believe there are many directions to continue offline TTI evaluation work. First, our models of visual salience and browsing, while validated on human preferences, could benefit from models customized for ideation tasks. Second, ideation and generation tends to be highly interactive, consisting of multiple turns where state can be carried across individual sessions and even between users. This complex interaction suggests that single turn metrics provide an incomplete picture of system effectiveness. Finally, our results may provide insight into how to design *online* evaluation metrics for TTI based on more elaborate behavior and interaction than found in classic information retrieval settings.

# References

[1] JM Ah-Pine, CM Cifarelli, SM Clinchant, GM Csurka, and Jean-Michel Renders. 2008. XRCE's Participation to ImageCLEF 2008. In *9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008)*.

[2] Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative Information Retrieval Evaluation. *arXiv preprint arXiv:2404.08137* (2024).

[3] Negar Arabzadeh and Charles LA Clarke. 2024. A Comparison of Methods for Evaluating Generative IR. *arXiv preprint arXiv:2404.04044* (2024).

[4] Negar Arabzadeh and Charles LA Clarke. 2024. Fr\'echet Distance for Offline Evaluation of Information Retrieval Systems with Sparse Labels. *arXiv preprint arXiv:2401.17543* (2024).

[5] Negar Arabzadeh, Oleksandra Kmet, Ben Carterette, Charles LA Clarke, Claudia Hauff, and Praveen Chandar. 2023. A is for Adele: An offline evaluation metric for instant search. In *ACM SIGIR International Conference on Theory of Information Retrieval*. 3–12.

[6] Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles LA Clarke. 2022. Shallow pooling for sparse labels. *Information Retrieval Journal* 25, 4 (2022), 365–385.

[7] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20041–20053.

[8] Shane Barratt and Rishi Sharma. 2018. A note on the inception score. *arXiv preprint arXiv:1801.01973* (2018).

[9] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. 2021. Demystifying MMD GANs. arXiv:1801.01401 [stat.ML]

[10] Ali Borji. 2022. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding* 215 (2022), 103329.

[11] Pia Borlund and Peter Ingwersen. 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation* 53, 3 (1997), 225–250.

[12] Sebastian Bruch, Shuguang Han, Mike Bendersky, and Marc Najork. 2020. A Stochastic Treatment of Learning to Rank Scoring Functions. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM 2020)*.

[13] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE TPAMI* 41, 3 (2018), 740–757.

[14] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*. 903–912.

[15] Ben Carterette and James Allan. 2007. Semiautomatic evaluation of retrieval systems using document similarities. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, New York, NY, USA, 873–876.

[16] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. 2022. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia* (2022), 1–1.

[17] Praveen Chandar, Fernando Diaz, and Brian St. Thomas. 2020. Beyond Accuracy: Grounding Evaluation Metrics for Human-Machine Learning Systems. In *Advances in Neural Information Processing Systems*.

[18] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 621–630.

[19] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*. 1–10.

[20] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491* (2022).

[21] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[22] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*. PMLR, 1931–1942.

[23] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. *CoRR* abs/2202.04053 (2022). arXiv:2202.04053 https://arxiv.org/abs/2202.04053

[24] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3043–3054.

[25] Min Jin Chong and David Forsyth. 2020. Effectively Unbiased FID and Inception Score and Where to Find Them. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6069–6078. https://doi.org/10.1109/CVPR42600.

[26] Min Jin Chong and David Forsyth. 2020. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6070–6079.

[27] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. 87–94.

[28] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero-and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. *arXiv preprint arXiv:2209.01390* (2022).

[29] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. 2023. The Infinite Index: Information Retrieval on Generative Text-To-Image Models. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (Austin, TX, USA) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 172–186. https://doi.org/10.1145/3576840.3578327

[30] Fernando Diaz, Ryen W. White, Georg Buscher, and Dan Liebling. 2013. Robust Models of Mouse Movement on Dynamic Web Search Results Pages. In *Proceedings of the 22nd ACM conference on Information and knowledge management (CIKM 2013)*. Association for Computing Machinery, New York, NY, USA, 1451–1460.

[31] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems* 34 (2021), 19822–19835.

[32] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Adversarial text-to-image synthesis: A review. *Neural Networks* 144 (2021), 187–209.

[33] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131* (2022).

[34] Yiqi Gao, Xinglin Hou, Yuanmeng Zhang, Tiezheng Ge, Yuning Jiang, and Peng Wang. 2022. Caponimage: Context-driven dense-captioning on image. *arXiv preprint arXiv:2204.12974* (2022).

[35] J. P. Guilford. 1950. Creativity. *American Psychologist* 5, 9 (1950), 444–454.

[36] Shengbo Guo and Scott Sanner. 2010. Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 833–834.

[37] 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. *Advances in neural information processing systems* 30 (2017).

[38] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2020. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1552–1565.

[39] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20406–20417.

[40] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023), 78723–78747.

[41] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. 2021. FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 576, 18 pages. https://doi.org/10.1145/3411764.3445093

[42] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1072–1080.

[43] Chandresh S Kanani, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Improving diversity and reducing redundancy in paragraph captions. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[44] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbozorgi. 2020. Creative Sketching Partner: An Analysis of Human-AI Co-Creativity. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 221–230. https://doi.org/10.1145/3377325.3377522

[45] Andruid Kerne, Andrew M. Webb, Steven M. Smith, Rhema Linder, Nic Lupfer, Yin Qu, Jon Moeller, and Sashikanth Damaraju. 2014. Using Metrics of Curation to Evaluate Information-Based Ideation. *ACM Trans. Comput.-Hum. Interact.* 21, 3 (jun 2014).

[46] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023),

36652–36663.

[47] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-Scale Text-to-Image Generation Models for Visual Artists' Creative Works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 919–933. https://doi.org/10.1145/3581641.3584078

[48] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI? Design Ideation with Cooperative Contextual Bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300863

[49] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 237–246.

[50] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.

[51] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. 2022. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18187–18196.

[52] Chunmian Lin, Lin Li, Wenting Luo, Kelvin CP Wang, and Jiangang Guo. 2019. Transfer learning based traffic sign recognition using inception-v3 model. *Periodica Polytechnica Transportation Engineering* 47, 3 (2019), 242–250.

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

[54] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou. 2018. An improved evaluation framework for generative adversarial networks. *arXiv preprint arXiv:1803.07474* (2018).

[55] Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G Baraniuk, and Michael J Henry. 2023. Evaluating generative networks using Gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 279–288.

[56] Loic Maisonnasse, Philippe Mulhem, Eric Gaussier, and Jean Pierre Chevallet. 2009. LIG at ImageCLEF 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 9*. Springer, 704–711.

[57] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (Dec. 2008), 2:1–2:27.

[58] William Morgan, Warren Greiff, and John Henderson. 2004. Direct Maximization of Average Precision by Hill-climbing, with a Comparison to a Maximum Entropy Approach. In *Proceedings of HLT-NAACL 2004: Short Papers (HLT-NAACL-Short '04)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 93–96.

[59] Vidhya Navalpakkam, Ravi Kumar, Lihong Li, and D. Sivakumar. 2012. Attention and Selection in Online Choice Tasks. In *User Modeling, Adaptation, and Personalization*, Judith Masthoff, Bamshad Mobasher, Michel C. Desmarais, and Roger Nkambou (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 200–211.

[60] Artem Obukhov and Mikhail Krasnyanskiy. 2020. Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance. In *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4*. Springer, 102–114.

[61] Jonas Oppenlaender. 2022. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*. 192–202.

[62] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[63] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14277–14286.

[64] Ville Paananen, Jonas Oppenlaender, and Aku Visuri. 2023. Using Text-to-Image Generation for Architectural Design Ideation. arXiv:2304.10182 [cs.HC]

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

[66] G. Parmar, R. Zhang, and J. Zhu. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 11400–11410. https://doi.org/10.1109/CVPR52688.2022.01112

[67] Nikita Pavlichenko and Dmitry Ustalov. 2022. Best Prompts for Text-to-Image Models and How to Find Them. *arXiv preprint arXiv:2209.11711* (2022).

[68] Vitali Petsiuk, Alexander E. Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A. Plummer, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, and Iddo Drori. 2022. Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark. In *NeurIPS 2022 Workshop on Human Evaluation of Generative Models (HEGM)*.

[69] Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics* 24, 2 (1975), 193–202.

[70] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *European conference on computer vision*. Springer, 647–664.

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

[72] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[73] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[74] Navyasri Reddy, Samyak Jain, Pradeep Yarlagadda, and Vineet Gandhi. 2020. Tidying Deep Saliency Prediction Architectures. In *IROS*.

[75] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*. 521–530.

[76] Cynthia L. Bennett Emily Denton Rida Qadri, Renee Shelby. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*.

[77] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[78] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=08Yk-n5l2Al

[79] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

[80] Jami J. Shah, Noe Vargas-Hernandez, and Steve M. Smith. 2003. Metrics for measuring ideation effectiveness. *Design Studies* 24, 2 (March 2003), 111–134.

[81] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. 2018. Chatpainter: Improving text to image generation using dialogue. *arXiv preprint arXiv:1802.08216* (2018).

[82] Chengyao Shen and Qi Zhao. 2014. Webpage Saliency. In *ECCV*. 33–46.

[83] Johanna M Silvennoinen and Jussi PP Jokinen. 2016. Appraisals of salient visual elements in web page design. *Advances in Human-Computer Interaction* 2016 (2016).

[84] Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. 2023. Balancing the Picture: Debiasing Vision-Language Datasets with Synthetic Contrast Sets. arXiv:2305.15407 [cs.CV]

[85] Gencer Sumbul, Sonali Nayak, and Begüm Demir. 2020. SD-RSIC: Summarization-driven deep remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing* 59, 8 (2020), 6922–6934.

[86] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[87] Roelof van Zwol, Vanessa Murdock, Lluis Garcia Pueyo, and Georgina Ramirez. 2008. Diversifying image search with user generated content. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. 67–74.

[88] Henriikka Vartiainen and Matti Tedre. 2023. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity* 34, 1 (2023), 1–21.

[89] Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. 2022. Too Many Relevants: Whither Cranfield Test Collections?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2970–2980. https://doi.org/10.1145/3477495.3531728

[90] Cheng Wang, Delei Chen, Lin Hao, Xuebo Liu, Yu Zeng, Jianwei Chen, and Guokai Zhang. 2019. Pulmonary image classification based on inception-v3 transfer learning model. *IEEE Access* 7 (2019), 146533–146541.

[91] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18359–18369.

[92] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *International Conference on Learning Representations*. https://openreview.net/forum?id=GUrhfTuf_3

[93] Norbert Wiener. 1960. Some Moral and Technical Consequences of Automation. *Science* 131, 3410 (1960), 1355–1358.

[94] Xiaoling Xia, Cui Xu, and Bing Nan. 2017. Inception-v3 for flower classification. In *2017 2nd international conference on image, vision and computing (ICIVC)*. IEEE, 783–787.

[95] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2024).

[96] Li-Chia Yang and Alexander Lerch. 2020. On the evaluation of generative models in music. *Neural Computing and Applications* 32, 9 (2020), 4773–4784.

[97] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2024. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems* 36 (2024).

[98] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=AFDcYJKhND

[99] Maia Zaharieva, Bogdan Ionescu, Alexandru-Lucian Gînsca, Rodrygo LT Santos, and Henning Müller. 2017. Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation.. In *MediaEval*.

[100] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 833–842.

[101] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. [n. d.]. LAFITE: Towards Language-Free Training for Text-to-Image Generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* ([n. d.]). https://par.nsf.gov/biblio/10351124