TIAM - A Metric for Evaluating Alignment in Text-to-Image Generation

Paul Grimal Hervé Le Borgne Olivier Ferret Julien Tourille

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

{paul.grimal, herve.le-borgne, olivier.ferret, julien.tourille}@cea.fr

Abstract

The progress in the generation of synthetic images has made it crucial to assess their quality. While several metrics have been proposed to assess the rendering of images, it is crucial for Text-to-Image (T2I) models, which generate images based on a prompt, to consider additional aspects such as to which extent the generated image matches the important content of the prompt. Moreover, although the generated images usually result from a random starting point, the influence of this one is generally not considered. In this article, we propose a new metric based on prompt templates to study the alignment between the content specified in the prompt and the corresponding generated images. It allows us to better characterize the alignment in terms of the type of the specified objects, their number, and their color. We conducted a study on several recent T2I models about various aspects. An additional interesting result we obtained with our approach is that image quality can vary drastically depending on the noise used as a seed for the images. We also quantify the influence of the number of concepts in the prompt, their order as well as their (color) attributes. Finally, our method allows us to identify some seeds that produce better images than others, opening novel directions of research on this understudied topic.

1. Introduction

The ability to generate synthetic images with neural models made significant advancements from the advent of the first GANs [12,26]. More recently diffusion-based models [1, 16, 23, 30, 31] have further pushed the boundaries of image synthesis by progressively denoising an initial noise to generate high-quality images. In parallel to these advances, the question of evaluating the quality of these synthetic images has always been a delicate issue and has become a research question in itself. To address this problem [34], several metrics were proposed [13,32,37] but they suffer from various limits [2, 4]. The most recent models are conditioned on textual image descriptions, allowing fine control of the output. It nevertheless adds a challenge to evaluate their outputs, namely to estimate to which extent the synthetic image generated corresponds to the textual description it was conditioned on.

Although Text-to-Image (T2I) models demonstrate strong semantic and compositional capabilities, achieving a visually pleasing image that aligns with the desired condition often requires the generation of multiple images to obtain a suitable one. A reliable generative model should exhibit alignment with the condition specified in the prompt, irrespective of the starting noise. To address and study the variability of results, we introduce a novel metric to assess the success rate of generative models according to a prompt, Text-Image Alignment Metric[1] (TIAM). The initial noise plays a crucial role in our metric, enabling us to investigate its impact. We show that certain initial noise configurations outperform others, suggesting the possibility of selecting them to get better synthetic images.

Recent research efforts [5, 9, 28, 31, 33] have shown that text-conditioned diffusion models suffer from three main issues related to the alignment between the expected content expressed in the textual prompt and the one actually generated in the image: (i) *catastrophic neglect*, where one or more elements described in the prompt are not generated or sometimes mixed, (ii) *attribute binding*, where attributes (*e.g*. color) are bound to the wrong entities, and (iii) *attribute leaking*, where attributes specified in the prompt are correctly bound but some other elements in the scene are also wrongly bound with this attribute (Fig. 1). With TIAM, we propose to analyze the success rate of generative models under the scope of catastrophic and attribute-binding issues. For the latter, we propose a solid method to evaluate color alignment with human perception.

To date, the investigation of the influence of some words and attributes remains largely understudied. Tang *et al*. [33] provide some insights by examining the impact of words on generated outcomes. Using TIAM, we provide further insights into the relationship between textual conditioning and generative results. Relying on prompt templates rather

[1] Source code: https://github.com/grimalPaul/TIAM

(a) Catastrophic neglect (b) Attribute binding (c) Attribute leaking

Figure 1. Images generated with the prompts "*a photo of a lion and a bear*", "*a photo of a blue cat and a yellow car*", and "*a photo of a red bus driving down the street*" generated with Stable diffusion v1.4. (a) The bear is missing, (b) the attributes are swapped, (c) the bus color (*red*) leaks on the wall

than natural prompts, TIAM allows quantifying the prompt-image alignment w.r.t syntactic aspects, in particular the importance of the position of the main entities in the prompt.

In summary, our main contributions are: (i) a new metric based on prompt templates to quantify automatically the performance of T2I models *in terms of prompt-image alignment*; (ii) an in-depth study of several diffusion models and *quantification* of their behavior relating to catastrophic neglect and attribute binding issues; (iii) a study on the influence of the initial diffusion noise (seed) for all these models. The main insights resulting from this study are (a) the alignment performance of most T2I models drops significantly with the number of objects specified in the prompt; (b) in practice there exists some seeds that systematically lead to better results; TIAM allows us to identify them and they still provide better results with objects that are out of our study domain; (c) most T2I models succeed in attributing color to one object but the performance drop with more objects.

2. Related Work

Text-to-image models Among the current state-of-the-art models, diffusion models [16] have demonstrated remarkable performance. These models introduce noise to images and learn to denoise the added degradation. During inference, the models iteratively denoise a noise sampling from a Gaussian distribution, resulting in a reconstructed image. The utilization of free classifier guidance [17] allows users to express their thoughts by writing a prompt using natural language. The diffusion process is then guided by the encoded text representation, leveraging foundational models like CLIP [25] or T5 [27]. It is important to note that the results are highly stochastic, often requiring the generation of multiple images with different Gaussian noise inputs to align with the user's desired texts and preferences. Among the well-known models in this field, notable mentions include Imagen [31], Dall-E 2 [28], Latent Diffusion Model (LDM) [30] and Ediff-I [1].

T2I Evaluation The evaluation of generated images often relies on metrics such as Inception Score (IS) [32] and Fréchet Inception Distance (FID) [13], which are commonly used to assess their quality and fidelity. However, these metrics do not capture the alignment between the provided condition(s) and the generated images. To address this issue, several methods have been proposed to measure whether the content of the generated images reflects the given conditions. One approach involves utilizing classification models [29] and object detection models (Semantic Object Accuracy, abbreviated as SOA) [15] to determine whether the generated images contain the specified objects or adhere to certain criteria. SOA uses actual prompts from COCO while our approach relies on prompt templates that allow a finer analysis of the influence of each element in the prompt. Another method specific to text-image alignment involves employing visual score similarity measures. It leverages contrastive models such as CLIP, which compute a similarity score between representations of images and text. Although CLIP has a great average semantic representation, it has poor compositional understanding [36], limiting a fine evaluation of text-image alignment. Our metric addresses this issue by leveraging a high-performance detector and by controlling the requested attributes in an explainable manner and aligning with human perception.

Recent efforts have introduced innovative methodologies to assess skills and measure biases of T2I models. Notably, Drawbench [31] presents a limited set of text prompts covering 11 skills to evaluate the models. Similarly, DALL-EVAL [6] proposes to evaluate three visual reasoning skills: the capacity to generate one object (object recognition), the capacity to generate the exact quantity number of the asked object (counting), and the capacity to place two objects (spatial reasoning). Additionally, they probe gender and skin bias in the model representation. Another contribution by Zhang *et al*. [38] delves into the gender depiction disparities enabling the study of potential stereotypes. Our methods focus on catastrophic neglect and attribute binding. We can see the catastrophic neglect approach as object recognition but by testing one or more different objects in the conditioning prompt. Moreover, to overcome results that are very sensitive to asked objects in prompt and evaluate the model's performance accurately systematically, we explore all various object and attribute combinations and we generate multiple images per prompt to enhance result robustness. Recognizing the impact of initial noises, we advocate for multiple seed testing. This approach offers a comprehensive evaluation, free from the limitations of specific labels, resulting in a more precise and less biased assessment. A closely related work [11] conducted concurrently with ours shares some conclusions, though it does not delve into the aspect of the starting noises.

3. Method

We propose a new method to measure the success rate of generative models w.r.t catastrophic neglect and/or attribute binding. First, we generate multiple prompts given a set of word labels and possible attributes. We then generate multiple images for each prompt and detect if the expected elements are present on the image, leading to the final score.

3.1. TIAM Text-Image Alignment Metric

Our approach is based on templates that are used to generate sets of prompts. We adopt a formalism inspired by the disentangled representation theory [14, 35]. The prompts contain N objects, each of which can be qualified by an attribute. Hence the object at position i in a prompt is a token belonging to the set $\mathcal{O}_i$ and qualified by an attribute in the set $\mathcal{A}_i$. For example, let us consider the template "*a photo of* $\texttt{det}(o_1,\ a_1)$ a_1 o_1 *and* $\texttt{det}(o_2,\ a_2)$ a_2 o_2" where $o_i \in \mathcal{O}_i$, $a_i \in \mathcal{A}_i$, and $\texttt{det}(o_i,\ a_i)$ is a determinant that depends on the object or attribute if present. If $\mathcal{O}_1 = \mathcal{O}_2 =${'car','bike','truck'} and $\mathcal{A}_1 = \mathcal{A}_2 =${'blue','green','red'}, then it can produce prompts such as "*a photo of a blue car and a red truck*" or "*a photo of a green bike and a blue car*".

For such a template t, the generic expression of the text-image alignment metric (TIAM) is defined as:

$$\mathop{\mathbb{E}}_{\substack{\chi \sim \mathcal{N}(0,I) \\ z \in \mathcal{Z}}} \left[f\left(G(\chi, t(z)), y(z)\right)\right] \quad \text{with:} \quad \mathcal{Z} = \prod_{i=1}^{N} (\mathcal{A}_i \times \mathcal{O}_i) \tag{1}$$

where the prompt is instanced from the template t and its "latent concepts" $z \in \mathcal{Z}$, $y(z)$ is the labels that relate to the expected content of the synthetic image generated by the model $G()$ (conditioned by the prompt) from a *seed* (using to generate an initial noise) χ. f is a scoring function that compares the ground truth $y(z)$ to the output of a model that detects objects or produces segmentation maps from the synthetic image, the resulting score being in $\{0, 1\}$, depending on whether the content matches the ground truth or not (see Section 3.3).

With the definition of Equation 1, a template can generate $\sum_{i=1}^{N} |\mathcal{A}_i|.|\mathcal{O}_i|$ different prompts, where $|.|$ is the cardinal of the set. However, we restrict in practice the inference of the template such that an object or an attribute can not appear twice in the prompt. In a simple case where the attribute and object sets are the same at each position, the number of unique prompts is given by:

Proposition 1: if $|\mathcal{O}| \geq N$, $|\mathcal{A}| \geq N$, and $\forall i \in [\![1, N]\!], \mathcal{A}_i = \mathcal{A}, \mathcal{O}_i = \mathcal{O}$, and $\forall (i,j) \in [\![1, N]\!]^2$ s.t $i < j$, we force $a_i \neq a_j$ and $o_i \neq o_j$, thus the number of unique prompts generated in the context of Equation 1 is $\frac{|\mathcal{O}|!|\mathcal{A}|!}{(|\mathcal{O}|-N)!(|\mathcal{A}|-N)!}$

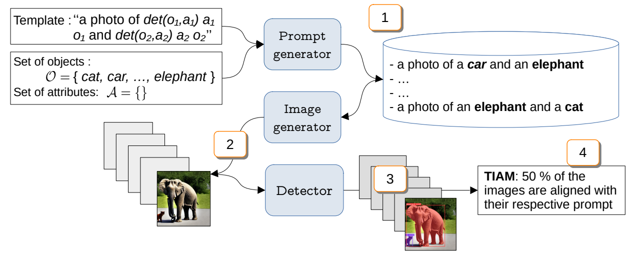

Figure 2. Overview of the evaluation pipeline. (1) Generate a dataset of prompts. (2) Generate $n \geq 16$ images per prompt. (3) Detect if the requested labels are present in the images. (4) Compute TIAM. In this example, we do not define attributes.

The proof is given in the Supp. Mat. (supplementary material). We also derive a more general case with possibly different attributes and objects at each position. The global evaluation process is summarized in Fig. 2.

3.2. Attribute definition

We study the color attribute in this paper, but our metric could be applied under the scope of another type of attributes (*e.g.*, size or texture). Selecting colors that are aligned with human perception is not trivial because of the infinity of possibilities. We base our choice on the work of Berlin and Kay [3]. They define eleven universal *basic colors* $\mathcal{C}$: *white, black, red, green, blue, brown, purple, pink, orange, yellow*, and *grey*. They asked individuals to select from an array of 329 colors (provided by the Munsell Colors Company) the chips that correspond to each *basic color* and to select the most typical examples. We use the results of the most typical examples of American English and convert the colors from the Munsell System to the CIELAB space. This one offers the advantage of a color distribution more aligned with human perception and a suitable space to compare color differences. We remarked that *brown* and *orange* were respectively too close to *black* and *white* and we removed them (more information provided in Supp. Mat.). We set $\mathcal{A} = \{red, green, blue, purple, pink, yellow\}$. However, to determine the actual colors of the pixels in the CIELAB space, we also consider *white* and *black*.

To define if $G()$ correctly assigns a_i and o_i, we use a segmentation model that delimits o_i if present on the image. We compare all the pixels inside the segmentation maps with our reference colors in the CIELAB space. If we detect at least a proportion of 40% of a_i, we consider a successful binding.

3.3. Main implementation details

The template t always includes the start mention "*a photo of*". Indeed, in our case, the detector is trained on real images and we suppose that it would therefore have more difficulty managing *e.g.* sketches or 3D renders. The general form of the templates is that given as an example above, by varying the syntactic context.

For the detection and segmentation tasks, we use YOLOv8 [18], a state-of-the-art model pre-trained on the 80 COCO labels [22]. We set the confidence threshold to 0.25 but conducted a study asserting that a higher threshold (0.4 to 0.8) does not change the relative order performance of the models. The results can be found in the Supp. Mat., with a detailed description of the labels used in each experiment and their combination within the generated prompts.

The scoring function f mainly focuses on true positives, considering the alignment as successful if we find at least one of each named object in the prompt. If the objects are characterized by an attribute, each object must be present with its correct attribute at least once in the image to be marked as a success. The criterion is therefore strict on the alignment aspect, requiring the presence of *all* the named objects and attributes. This approach differs from CLIP scores, which provide a general trend without reflecting a comprehensible explanation of the results. It nevertheless does not penalize the presence of other objects (false positive) since it is not prohibitive in a generative context. Indeed, if the segmentation masks of two different objects overlap with an IoU value of 0.95 or higher, we remove both objects before calculating the score with the remaining detected objects.

4. Study

We conduct an in-depth study to characterize and quantify the limits of several diffusion models, which differ by the text-conditioning method, the architecture, or the inference process. We consider Stable Diffusion v1-4 (SD 1.4) and Stable Diffusion v2 (SD 2) based on the LDM model [30] that uses a fixed pre-trained text encoder CLIP [25], which is based on an auto-regressive architecture. Both versions are trained with the same variational autoencoder (VAE), but two different U-nets (the part that learns to denoise image). The CLIP models that guide the U-nets also differ between SD 1.4 and SD 2. In addition, we evaluate the two SD models with Attend-and-Excite (A&E) [5], an optimized inference process that uses the cross-attention maps to attend the subject tokens in the prompt. We follow the authors' implementation and only attend to the token of object o_i, even if an attribute characterizes the object. We also consider an unCLIP model [20, 28] conditioned with a CLIP image prior and a CLIP text embedding. Finally, we study DeepFloyd IF (IF), a cascade diffusion inspired by [31], conditioned using a T5 XXL [27] text encoder. Other details are reported in the Supp. Mat.

4.1. Preliminary: performance drop with 2 objects

We report a first experiment to illustrate the general framework of our study. We consider various labels from COCO (the exhaustive list is in Supp. Mat.), leading to a set of $|\mathcal{O}| = 24$ object labels. For all possible pairs of labels (o_i, o_j) with $o_i \neq o_j$, we make the prompt "a photo of o_i and o_j" (managing the determinant as expected) and generate 64 synthetic images by changing the random seed χ and estimate the alignment of the prompt with each image using TIAM. As a reference, we also conduct the same experiment with 64 images generated from the simpler prompt "a photo of o_i".

Model	1 object	2 objects
SD 1.4	0.98	0.41
SD 1.4 A&E	0.96	0.64
SD 2	0.99	0.61
SD 2 A&E	0.98	0.65
unCLIP	0.95	0.50
IF	0.99	0.62

Table 1. TIAM according to the number of objects in the prompt, for all 6 generative models considered in the study.

We report the results in Tab. 1 for all the models considered. The models consistently succeed in generating images with a single object, with a score above 0.95. However, they struggle to generate simultaneously two objects, with at most 66% of images correctly generated, while the prompts are minimally simple. In the recent literature, the T2I models usually exhibit a good quantitative rendering score, *e.g*. in terms of inception score or FID, which our experience does not call into question. In the vein of Hinz *et al*. [15], this experiment only shows that there is a problem with alignment between the content of the generated image and the prompt. More precisely than [15], our experiment specifically quantifies to which extent this drop in alignment performance is due to the presence of multiple concepts in the prompt. For the prompts with two objects, we report in Tab. 2 the number of times the objects are actually accurately generated with regard to their position in the prompt (the score in Tab. 1 requires the presence of both objects to be valid). We observe a tendency for the first o_1 to be more prevalent than the second o_2 across all models, a phenomenon that is further investigated in Section 4.3.

We notice that A&E inference globally improves the score with a large margin for SD v1.4. From Tab. 2, it seems that the overperformance of A&E is due to its ability to enhance the occurrence rate of o_2 by enforcing the minimum excitation of cross-attention maps.

4.2. Importance of the random seed

We consider the same 24 labels as in Section 4.1 and the prompt "a photo of o_i and o_j". We also generate 64 synthetic images by changing the random seed χ and compute $f\left(G(\chi, t(z)), y(z)\right)$ for each prompt and the corresponding

Model	o_1	o_2
SD 1.4	0.80	0.60
SD 1.4 A&E	0.85	0.75
SD 2	0.83	0.78
SD 2 A&E	0.84	0.80
unCLIP	0.77	0.71
IF	0.86	0.76

Table 2. Proportion of appearance per order in the prompt. o_1 and o_2 refer to the position in the template.

image. However, this time, the 64 seeds are the same for all the (o_i, o_j)[2] and we aggregate the performance *per seed*, for all possible prompts and images, leading to $24 \times 23 = 552$ estimations of text-image alignment per seed.

The results in Fig. 3 show that, for all 6 models, the performance varies a lot with regard to the random seed. These results, which to the best of our knowledge have never been identified in the literature of the T2I generative models, may seem surprising. With such models, one usually expects that all seeds have the same chance to generate the specified elements. Actually, the recent tendency rather consists of engineering finely the prompt [24] to optimize the output of T2I models. On the opposite of this tendency, our method opens up new perspectives toward a complementary optimization based on the choice of "performing seeds" (see Section 4.6).

Overall, there exists a significant disparity between the best and worst seeds. When considering the interquartile ranges and the min-max range through the box-plots of Fig. 3, it appears that the difference in average performance between models is much less significant than their own inner difference due to the variation of the seeds.

Obviously, enhancing a dependence "to the seed" is convenient from a practical point of view but wrong in all strictness since the initial noises are drawn from a Gaussian at inference. Being more rigorous requires reminding the diffusion models training process [16]. It consists of learning the reverse process of a fixed Markov chain of length T with models that can be interpreted as an equally weighted sequence of denoising autoencoders. The latters are trained to denoise their input, considered as a noisy version of an input training image I. In the case of latent diffusion models [30], the process is embedded into the latent space of a VAE (encoder $\mathcal{E}$ + decoder $\mathcal{D}$), such that the autoencoders are U-net networks $\epsilon_\theta(x_t, t)$, $t = 1 \dots T$ that denoise the latent code x_t by minimizing the loss $\mathbb{E}_{\mathcal{E}(I),\epsilon\sim\mathcal{N}(0,1)}\left[||\epsilon - \epsilon_\theta(x_t, t)||_2^2\right]$. During inference, the process starts from a latent code $\chi_T \sim \mathcal{N}(0, 1)$, denoises it with the U-net ϵ_θ to get the final latent code χ_0, then obtains the synthetic image with VAE decoder as $\mathcal{D}(\chi_0)$. This

[2] Actually, it was the case in Section 4.1 for the sake of comparison and coherency.

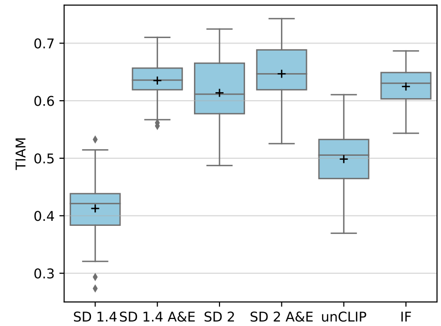

Figure 3. TIAM aggregate per seed for 64 seeds. We show that some starting noise tends to not be converted to an image with two entities regardless of the entities. "+" shows the mean.

last is thus not just dependent on the random seed χ_T but on the U-net parameter θ as well. These parameters are learned to denoise the latent codes at any step of the diffusion since the reparametrization trick [16] allows to express the final code directly as:

$$x_0 = \frac{x_t - \sqrt{1-\bar{\alpha_t}}\epsilon}{\sqrt{\bar{\alpha_t}}} \quad (2)$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and the α_t depends on the variance of the diffusion process noise (see [16] and appendix B of [30] for details). Ideally, $\bar{\alpha_T}$ is close to one, such that a synthetic image should result from an almost perfect Gaussian noise. However, in practice the process is not perfect; thus the final χ_0 obtained for a given $\chi_T \sim \mathcal{N}(0, 1)$ at inference time is only an approximation of the "ideal χ_0" that could be expected with a perfect optimization [7]. Moreover, since each model is trained independently, its latent space is structured in the same vein and the "path" of the χ_t in the latent space may strongly differ from one model to another. Hence, starting from the same $\chi_T \sim \mathcal{N}(0, 1)$, they result in quite different χ_0 while trying to tend to an ideal one. As a consequence, the "good" and "bad" seeds are specific to each model.

The results depend on the seed, indicating that based on the prompt's specifications, it appears possible to identify initial noises that are more likely to exhibit multiple objects. It highlights the ongoing need for further advancements in reducing reliance on latent variables. To obtain a robust score and avoid the possible impact of the high variability of the seed's success rate, we define a minimum of images to generate per prompt. We establish it is adequate to generate 32 images per prompt to obtain a robust score. The corresponding experiment is reported in the Supp. Mat..

4.3. Catastrophic Neglect

We study the behavior of the models as the number of object sets increases and the role of the order of objects in the prompt. We set $\mathcal{O} = \{$*car, refrigerator, giraffe, elephant, zebra*$\}$ and compute TIAM for prompts containing from 1 to 4 objects. We design 4 templates, one per number of objects in the prompt, and generate 32 images for each prompt, leading to 160 alignment values, 640 values, 1,920 values, and 3,820 values respectively with one, two, three, and four objects in the prompt (for each model). As shown in Fig. 4, the models fail to consistently generate outputs when prompted with more than two objects. Even with the A&E mechanism, generating four objects remains nearly impossible. SD 2 and IF demonstrate relatively better performance, but the improvement is not substantial.

When examining the occurrence of different objects, similar trends can be observed as in the preliminary experiment (Tab. 2). Specifically, the initial objects in the template tend to appear more frequently than objects inserted subsequently. The results for the template containing four objects are presented in Fig. 5 (results for two and three objects in Supp. Mat.). This reinforces the observation that the concept that is expressed earlier in the prompt has more chances to appear in the final image.

For SD, conditioned by CLIP text-encoders, the decreasing trend in the occurrence of objects as their position becomes more distant in the prompt may be partly due to the auto-regressive nature of the encoder. During self-attention, tokens only receive context from the elements to their left (beginning of the prompt). Tokens are thus devoid of the contexts of subsequent objects, while these last carry the context of the earlier words. Hence, the U-net model receives a more ambiguous signal from distant tokens. The explanation is less clear with the T5 encoder, where all words have access to each other during self-attention. Used in IF, one can see that the third and fourth objects have the same chance to appear, but it is significantly less than the object in the second position, itself below the first one expressed in the prompt. We hypothesize that during cross-attention, the models learned to give more importance to tokens with earlier positions due to the training data, which typically places important elements related to the image at the beginning of the caption. Asserting this explanation would nevertheless require significant work to analyze the original training dataset used to pre-train the six generative models, which is out of the scope of this paper.

Finally, we investigated the impact of semantic relationships between objects within the prompt on the ability of the model to generate both of them. Considering a template with two objects, we hypothesized that the T2I model would fail more often to represent in the same image two objects that are semantically linked. We considered 28 COCO labels from three macro-classes (*vehicles, animals*, and *foods*) and generated images using a template with 2 objects. From the resulting TIAM scores, we derived a dissimilarity metric between all objects (see Supp. Mat.) and projected all the

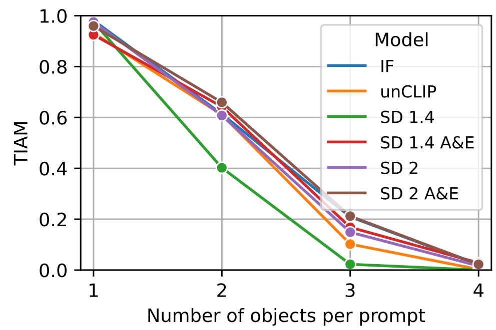

Figure 4. TIAM with 1 to 4 objects per prompt.

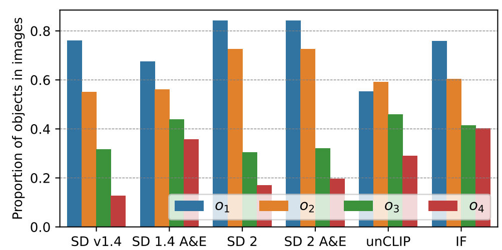

Figure 5. The proportion of occurrences of each object, based on its position in the prompt (template with 4 objects).

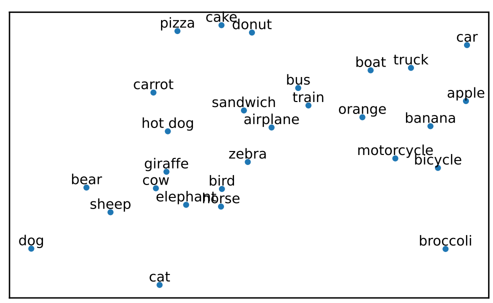

Figure 6. MDS on the objects score dissimilarity for SD 1.4.

labels with Multidimensional Scaling (MDS). The resulting projections (Fig. 6 for SD 1.4) can be interpreted such that the closer two labels are, the more challenging it becomes for the model to represent them together. Across all models, we observe labels from the same class being clustered together, particularly for *animals* and *vehicles*. This suggests the presence of semantic proximity. We measured the semantic distance between both named objects with various methods (such as Wu-Palmer, Cosine similarity of CLIP text embeddings, or even attention's key-value representation) for SD 1.4 and SD 2 but the correlation with the TIAM score was only slightly negative. This indicates that semantic linking has probably either a small or indirect link to the alignment performance, but in any case, further research is needed to clarify this point.

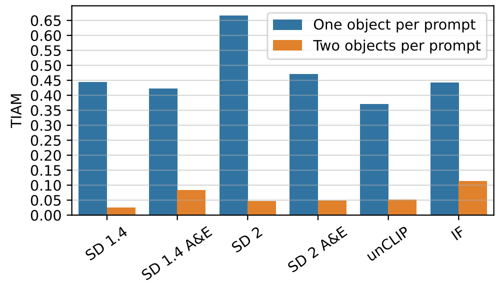

Figure 7. TIAM computed with object and color ground truth, for one and two colored objects per prompt.

	1 object		2 objects	
Attribute	✗	✓	✗	✓
SD 1.4	0.97	0.95	0.40	0.18
SD 1.4 A&E	0.93	0.94	0.64	0.42
SD 2	0.97	0.97	0.61	0.26
SD 2 A&E	0.96	0.96	0.66	0.32
unCLIP	0.93	0.89	0.61	0.36
IF	0.98	0.90	0.61	0.49

Table 3. TIAM computed with object(s) ground truth only, with prompts containing attributes or not.

4.4. Attribute binding

In this section, we characterize the capacity of models to apply attributes to objects. Let us consider 2 prompts with respectively one and two objects using the same set $\mathcal{O} = \{$*car, refrigerator, giraffe, elephant, zebra*$\}$ and $\mathcal{A}$ the set of colors defined in Section 3.2.

We start by computing TIAM regardless of the correctness of the attribute and report the results in Tab. 3. Without attributes (✗) the scores are close to those of Tab. 1 and adding attributes (✓) has a limited impact when one object only is requested. However, it significantly impacts the performance when two objects are present in the prompt.

When TIAM is estimated by taking into account both the object and color ground truth, we observe that the models fail to assign the appropriate colors to the objects, even when these objects are present in the image (Fig. 7). For instance, in the case of SD 1.4, objects are detected correctly in 95% of single-object prompts, but only approximately 45% of them have the requested colors.

In Fig. 8, we report the attribution scores per color and object to analyze deeper the attributing ability of the models. We see that models fail to generalize the binding to objects when it is uncommon to observe them in that particular color. As expected, it is indeed easier to assign colors to cars and refrigerators compared to animals. The models have likely been trained on photos of cars and refrigerators in various colors during training as it is more common to come across a green car than a green giraffe.

Following the same approach as in Section 4.3, we also analyze the results concerning the position i of attribute a_i and object o_i in the prompt involving 2 objects. In that case, the models succeed by a large margin to generate and bind the first object. However, knowing that the first object is more often generated, we compute a *binding success rate*, which is the score of correctly attributed objects among the correctly generated objects (Fig. 9). The o_2 objects, however, continue to be less attributed. This reinforces the finding that the first object in the prompt has a greater influence on the final generation. In the Supp. Mat., we report the results of the *binding success rate* differentiated by colors for attributes in the first position and attributes in the second position. We observed that the models face greater difficulty in assigning green and blue colors when two objects are involved (parallel with a single label case). It is worth noting that IF performs better than other models.

4.5. Comparison to human, CLIP, BLIP

We randomly chose 32 prompt-image pairs and asked 57 humans to assess the content alignment. Their reliability of agreement was 0.73 in terms of Fleiss'kappa [10], which can be considered as "Substantial" [19]. Half of the prompts had two objects and the 16 others had one colored object. The Pearson correlation of humans and TIAM was 0.82 ($p < 10^{-8}$). We compared TIAM to two other automatic methods based on CLIP [25] and BLIP [21], which had respectively a correlation of 0.47 ($p < 10^{-2}$) and 0.67 ($p < 10^{-4}$) with humans. As detailed in the Supp. Mat., TIAM had a better alignment with humans both for images with two objects and with one colored object.

4.6. Toward noise mining?

To highlight the importance of noise performance, we selected seeds based on their TIAM score. In Fig. 10, we present qualitative results using some of the worst and best seeds found with our approach, as well as prompts with objects and attributes that were not considered in our study. Both with SD 2 and IF, the two objects are better represented in the image resulting from the "good" seed, showing that our method has the ability to find seeds that generalize to objects out of the domain of objects used to determine the best seeds. We conducted the same seed selection process for both our worst and best seeds using colors in the prompt (Fig. 11). Once again the images resulting from the best seeds we identified with TIAM better reflect the prompt. With the "bad" seed, SD 2 suffers both from attribute leaking or binding (the "blue moon" is yellow, and the "red lion" is blue), and for both models, the images resulting from the "good" seed are more aligned. This emphasizes the dependency on the noise present at the beginning of the prompt, which remains a prominent factor of performance.

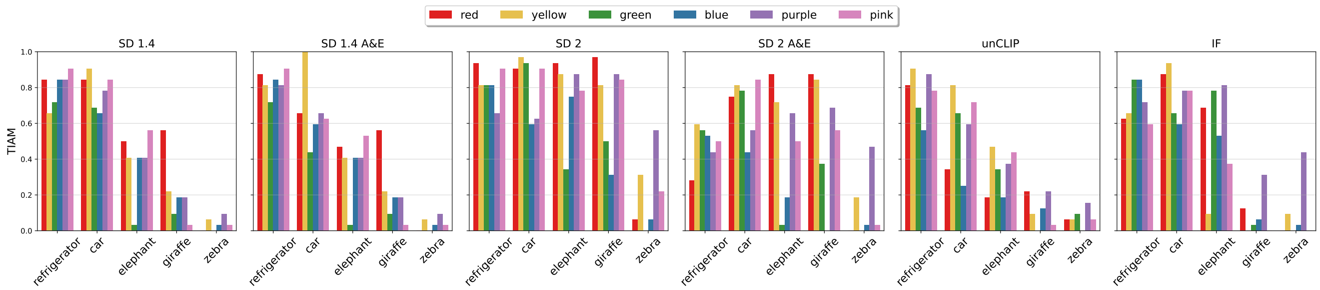

Figure 8. TIAM per color and object.

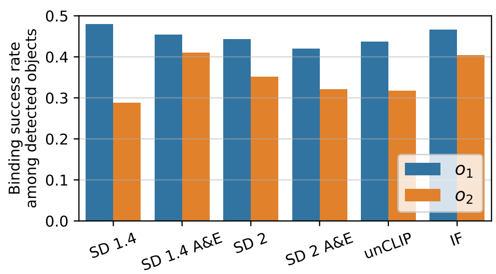

Figure 9. Success rate of color attribution w.r.t the detected objects.

Figure 10. Generation of images with the same prompt and a "bad" (left) of a "good" (right) random seed. *top*: SD 2 "a sketch of a tiger and a surfboard, 4k, 8k, ghibli" seed 23, 11. *bottom*: IF "a photo of space tiger and a rocket" seed 41, 9.

5. Conclusion, Limits and Perspective

We proposed a new metric to automatically quantify the performance of T2I models in terms of prompt-image alignment. Contrary to previous research efforts, it is based on prompt templates, that allow a finer analysis with regard to the syntax of the prompt. Hence, we showed that the alignment performance of most T2I models drops significantly with the number of objects specified in the prompt and that effect is even more critical for color attribution. Extending

Figure 11. Generation of images with the same prompt and a "bad" (left) of a "good" (right) random seed. *top*: SD 2 "a drawing of a red lion and a blue moon, pop art, 4k, highly detailed" seed 27, 17. *IF* "a photo of a purple frog and a rainbow piano", seed 19, 24.

TIAM with more objects and attributes would result in the need to generate an exponential number of prompts, thus becoming cumbersome in practice. However, we show in Section 12 of the Supp. Mat. that TIAM can be reliably estimated from a set of ≈ 300 prompts. Our metric also allows us to study the influence of the input seed at inference. We showed that there exist some seeds that systematically result in better output images than others and that it generalizes to objects out of the set used to determine them. It draws possible future research toward the mining of such "good seeds", similarly to some studies for text models [8], as a complementary activity to prompt engineering to optimize the outputs of T2I models. For a comprehensive analysis of T2I performance, TIAM should be combined with other metrics reflecting other aspects than prompt-image alignment, such as [6, 13, 31, 32, 38].

Acknowledgement This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011014009 made by GENCI. This publication relied on the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv 2211.01324*, 2023. 1, 2

[2] Shane Barratt and Rishi Sharma. A note on the inception score. In *Proc. ICML 2018 Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018. 1

[3] Brent Berlin and Paul Kay. *Basic Color Terms: Their Universality and Evolution*. University of California Press, Los Angeles, 1969. 3

[4] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 1

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4), jul 2023. 1, 4

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv 2202.04053*, 2022. 2, 8

[7] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. 5

[8] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020. 8

[9] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[10] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378—-382, 1971. 7

[11] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. 1

[13] 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 *Advances in Neural Information Processing Systems*, page 6629–6640, Dec 2017. 1, 2, 8

[14] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv 1812.02230*, 2018. 3

[15] 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(3):1552–1565, mar 2022. 2, 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1, 2, 5

[17] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv 2207.12598*, 2022. 2

[18] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Yolo by ultralytics, 2023. 4

[19] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174, 1977. 7

[20] Donghoon Lee, Jiseob Kim, Jisu Choi, Jongmin Kim, Minwoo Byeon, Woonhyuk Baek, and Saehoon Kim. Karlo-v1.0.alpha on coyo-100m and cc15m. https://github.com/kakaobrain/karlo, 2022. 4

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 7

[22] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 4

[23] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022. 1

[24] Jonas Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation. *arXiv 2204.13988*, 2023. 5

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 2, 4, 7

[26] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016. 1

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv 1910.10683*, 2020. 2, 4

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv 2204.06125*, 2022. 1, 2, 4

[29] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In *Advances in Neural Information Processing Systems*, 2019. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1, 2, 4, 5

[31] 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. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 2, 4, 8

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 1, 2, 8

[33] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, Toronto, Canada, July 2023. Association for Computational Linguistics. 1

[34] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, Apr 2016. 1

[35] Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: Compositional structures in vision-language models. In *ICCV 2023*, 2023. 3

[36] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. 2

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1

[38] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *arXiv 2302.03675*, 2023. 2, 8