arXiv:2212.10015v2 [cs.CV] 6 Jun 2023

# Benchmarking Spatial Relationships in Text-to-Image Generation

Tejas Gokhale [1*] Hamid Palangi [2] Besmira Nushi [2] Vibhav Vineet [2]
Eric Horvitz [2] Ece Kamar [2] Chitta Baral [1] Yezhou Yang [1]
[1]Arizona State University [2] Microsoft Research

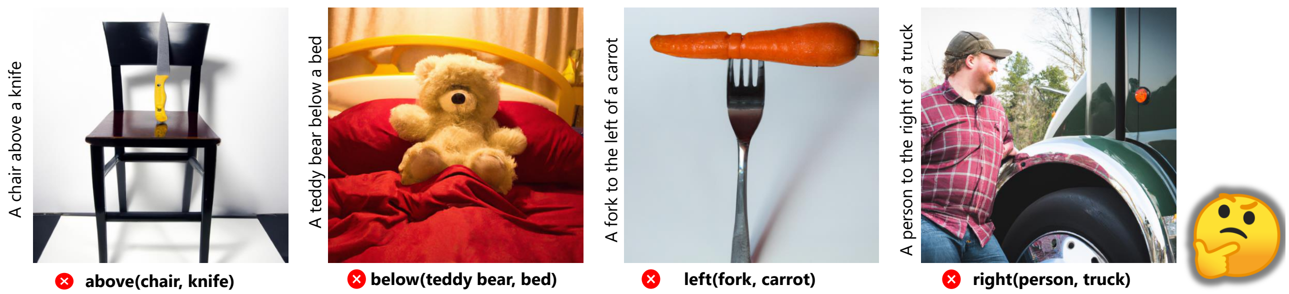



Figure 1: We benchmark T2I models on their competency with generating appropriate spatial relationships in their visual renderings. Although text inputs may explicitly mention these spatial relationships, T2I models lack such spatial understanding.

## Abstract

*Spatial understanding is a fundamental aspect of computer vision and integral for human-level reasoning about images, making it an important component for grounded language understanding. While recent text-to-image synthesis (T2I) models have shown unprecedented improvements in photorealism, it is unclear whether they have reliable spatial understanding capabilities. We investigate the ability of T2I models to generate correct spatial relationships among objects and present VISOR, an evaluation metric that captures how accurately the spatial relationship described in text is generated in the image. To benchmark existing models, we introduce a dataset, $SR_{2D}$, that contains sentences describing two objects and the spatial relationship between them. We construct an automated evaluation pipeline to recognize objects and their spatial relationships, and employ it in a large-scale evaluation of T2I models. Our experiments reveal a surprising finding that, although state-of-the-art T2I models exhibit high image quality, they are severely limited in their ability to generate multiple objects or the specified spatial relations between them. Our analyses demonstrate several biases and artifacts of T2I models such as the difficulty with generating multiple objects, a bias towards generating the first object mentioned, spatially inconsistent outputs for equivalent relationships, and a correlation between object co-occurrence and spatial understanding capabilities. We conduct a human study that shows the alignment between VISOR and human judgement about spatial understanding. We offer the $SR_{2D}$ dataset and the VISOR metric to the community in support of T2I reasoning research.* [1]

## 1. Introduction

Text to image synthesis (T2I), has advanced rapidly with capabilities for generating high-definition images in response to text prompts. Models are being used as tools for art, graphic design, and image editing. The power of T2I models for generating photorealistic objects and scenes is well-known. We less understand the ability of the models to faithfully render spatial relationships in its compositions.

We pursue the question: Do T2I models have the ability to render the spatial relationships among objects that are specified in text prompts? Fig. 1 illustrates images generated by a state-of-the-art model (DALLE-v2 [36]) for sentences that contain a spatial relationship between two objects. In these examples, although both objects mentioned in the text are generated, the specified spatial relationship is not rendered.

Spatial relations and larger scene geometries are integral aspects of computer vision. Rendering and reasoning about these relationships is crucial for many applications such as language-guided navigation and object manipulation [3, 30, 32]. A lack of spatial understanding by T2I models can be frustrating to creators seeking to render specific configurations of objects. The assertion of spatial relationship is common in natural communication among humans and poor capabilities in this realm will rapidly come to the fore in navigational and instructional applications.

Prior work on evaluation metrics for T2I models have focused on photorealism [41, 17], object accuracy [18], and image-text vector similarity (via CLIP [16], retrieval [44], and captioning [19]). We find that these metrics are insensi-

[1]Data and code: https://github.com/microsoft/VISOR.

tive to errors with generating spatial relationships (Sec. 5.1). This finding highlights the need for a metric to quantify competencies and progress in spatial understanding. We develop an automated evaluation pipeline that employs computer vision to recognize objects and their spatial relationships, and harness this pipeline to conduct a large-scale evaluation of the spatial understanding capabilities of T2I models. We create the "$\text{SR}_{\text{2D}}$" dataset, containing 25,280 sentences describing two-dimensional spatial relationships (*left/right/above/below*) between pairs of commonly occurring objects from MS-COCO [25], as shown in Table 1. We study several state-of-the-art models: GLIDE [33], DALLE-mini [10], CogView2 [11], DALLE-v2 [36], Stable Diffusion [39], and Composable Diffusion [27]. For each model we generate and evaluate four images per $\text{SR}_{\text{2D}}$ example, i.e., a large-scale study of 101,120 images per model. Our study makes significant advances to evaluation of T2I reasoning capabilities since we evaluate photorealistic images rather than synthetic objects on solid background.

We introduce a new evaluation metric we refer to as VISOR (for **v**erify**i**ng **s**patial **o**bject **r**elationships), to compare the spatial understanding abilities of T2I models. We define three variants of the metric: (1) VISOR: verifies spatial correctness for each image w.r.t. its text input, (2) $\text{VISOR}_{\text{n}}$: consider whether at least $n$ of the multiple generated images for each text input are spatially correct, (3) $\text{VISOR}_{\text{cond}}$: verifies the spatial correctness in images, conditioned on both objects being generated by the model. While VISOR provides a macro-perspective on the performance gap in the spatial capabilities of T2I models, $\text{VISOR}_{\text{n}}$ reflects the practical value of the model to users who can select one of many images generated by the model. The conditional formulation $\text{VISOR}_{\text{cond}}$ disentangles two capabilities: (i) the generation of multiple objects and (ii) generation of correct spatial relationships between the rendered objects. We conduct a human study on Amazon Mechanical Turk and find that the VISOR metric is correlated with human judgment.

Our experiments reveal several interesting findings. First, we find that all existing models are significantly worse at generating two objects as compared to their capability to render single objects. While previous work shows exceptional zero-shot compositionality of colors, styles, and attributes [36, 40, 46], we found challenges with compositionality for multiple objects. Second, we find poor spatial understanding: even in cases where both objects are generated, models tend to ignore spatial relationships specified in language. VISOR scores for all models show that even the best model in our benchmark generates correct spatial relationships on less than 40% of test cases. When we consider a strict metric ($\text{VISOR}_4$) that requires that all generated images for text prompts to have correct spatial relationships, the best model (DALLE-v2) achieves the goal in 7.49% cases. Third, we discover several biases in T2I models: to generate only the first object mentioned in the text and ignoring the second, to show better performance on commonly occurring object pairs, to have a tendency to merge two objects into one, and to have inconsistent outputs for equivalent text inputs.

| **A** | **B** | **R** | **Text** |
|---|---|---|---|
| microwave | sink | left | A microwave to the left of a sink |
| elephant | cat | right | An elephant to the right of a cat |
| donut | airplane | above | A donut above an airplane |
| suitcase | chair | below | A suitcase below a chair |
| keyboard | bench | left | A keyboard to the left of a bench |
| bed | bear | right | A bed to the right of a bear |
| potted plant | fire hydrant | above | A potted plant above a fire hydrant |
| person | umbrella | below | A person below an umbrella |

Table 1: Examples text inputs from the $\text{SR}_{\text{2D}}$ dataset for a pair of objects (A, B) and relationship R between them.

To summarize, our contributions are as follows:

- We introduce a metric called VISOR to quantify spatial reasoning performance. VISOR can be used off-the-shelf with any text-to-image model, disentangles correctness of object generation with the ability of spatial understanding.
- We construct and make available a large-scale dataset: $\text{SR}_{\text{2D}}$, which contains sentences that describe spatial relationships between a pair of 80 commonly occurring objects along with linguistic variations.
- With $\text{SR}_{\text{2D}}$, we conduct a large-scale benchmarking of state-of-the-art T2I models with automated and human evaluation of spatial reasoning abilities of state-of-the-art T2I models using the VISOR metric. We find that although existing T2I models have improved photorealism, they lack spatial and relational understanding with multiple objects, and indicate several biases.

## 2. Related Work

**Text-to-Image Synthesis.** Earlier work [38, 47] trained and evaluated models on human-labeled datasets [43, 34, 25]. Recent work on T2I has focused on zero-shot capabilities by taking advantage of implicit knowledge from pretrained language models and V+L models like CLIP, and the diffusion technique to train on large-scale web data.

**Biases in Vision+Language models** have been studied from a linguistic perspective, such as question-answer priors in VQA [1, 22], gender bias in captioning [15, 48], shortcut effects in commonsense reasoning [45], and failure modes in logic-based VQA [37, 14, 13]. The difficulty of spatial understanding has been studied for visual grounding [28], image-text matching [26], VQA [21, 20], and navigation [7].

**Human Study about Relational Understanding.** Conwell *et al.* [9] conducted a human study (1350 images) of DALLE-v2 on a set of eight physical relations and seven action-based relations between 12 object categories. Our human study is significantly larger in scale, considers diverse text inputs, several state of the art models, and establishes an alignment with the automated VISOR metric.

**Empirical Evaluation of Visual Reasoning Skills.** DALL-Eval [8] evaluates reasoning skills of T2I models trained and tested on a synthetically generated dataset PAINTSKILLS with black backgrounds and 21 rendered object categories. In our work, we instead focus on the evaluation of photorealistic and open-domain images with commonly occurring real-world objects and backgrounds on a large scale. Most importantly, we devise a new human-aligned metric (VISOR) that disentangles object accuracy from spatial understanding to accordingly measure progress in spatial reasoning despite the model's capabilities in object generation.

**Other Failure Modes of T2I Models.** Preliminary stress-testing of DALLE-v2 [29] (14 prompts), [23] (40 prompts) and [40] (200 prompts) illustrated anecdotal failures of the model in terms of compositionality, grammar, binding, and negation. However, since these studies rely on human judgment, there is a need for automated evaluation techniques for comparing the reasoning abilities of T2I models. Our paper fills this gap with the automated VISOR metric for spatial relationships and the large-scale $SR_{2D}$ dataset.

## 3. Spatial Relationships Challenge Dataset

**Predicate Generation.** Our goal is to collect a set of sentences that describe spatial relationships between two objects. Let $\mathcal{C}$ be the set of object categories. Let $\mathcal{R}$ be the set of spatial relationships between objects. In this paper, we focus on two-dimensional relationships, i.e. $\mathcal{R} = \{left, right, above, below\}$, and 80 object categories derived from the MS-COCO dataset [25]. Then, for every $A \in \mathcal{C}$, $B \in \mathcal{C}$, and $R \in \mathcal{R}$, let the predicate $R(A, B)$ indicate that the spatial relationship $R$ exists between object $A$ and object $B$. For example `left(cat,dog)` describes a scene where a cat is to the left of a dog. For each pair, we construct 8 types of spatial relationships as shown below:

| |
|---|
| *left(A, B), right(A, B), above(A, B), below(A, B)* |
| *left(B, A), right(B, A), above(B, A), below(B, A)* |

**Sentence Generation.** For each predicate $R(A, B)$, we convert it into a template `<A> <R> <B>` and paraphrase it into natural language. Appropriate articles *"a"/"an"* are prepended to object names `A` and `B`, to obtain four templates:

| |
|---|
| A/an `<A>` *to the left of* a/an `<B>` |
| A/an `<A>` *to the right of* a/an `<B>` |
| A/an `<A>` *above* a/an `<B>` |
| A/an `<A>` *below* a/an `<B>` |

The template-based procedure has several advantages. First, it avoids linguistic ambiguity, subjectivity, and grammatical errors. Second, it is extensible to new object categories and additional spatial relationships. While we focus on two-dimensional relationships in this paper, our templates

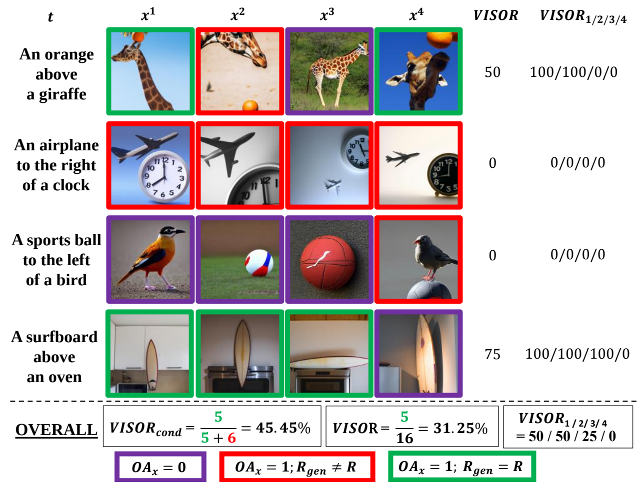



Figure 2: Examples illustrating the intuition behind OA, VISOR, $\text{VISOR}_{\text{cond}}$, and $\text{VISOR}_{1/2/3/4}$. **Purple box**: cases where one or both objects are not generated; **Red box**: both objects are generated but with a wrong spatial relationship; **Green box**: successful cases.

can be extended for generating test inputs for studying more complex spatial relationships and geometric features of objects, as we discuss in Sec. 7.

**Dataset Statistics.** We use $|\mathcal{C}| = 80$ object categories from MS-COCO and therefore obtain $\binom{80}{2} = 3160$ unique combinations of object pairs (A, B). For each pair, we construct 8 types of spatial relationships listed above, which leads to a total of 3,160$\times$ 8 = 25,280 predicates. The $SR_{2D}$ dataset contains 25,280 text prompts, uniformly distributed across 80 COCO object categories, with each object being found in 632 prompts. Tab. 1 lists a few illustrative examples.

## 4. VISOR Metric

We propose VISOR as an automated metric for quantifying spatial understanding abilities of text-to-image models.

**Definition 1 (Object Accuracy)** *Let $h$ be an oracle function that returns a set of detected objects in image $x$ from set $\mathcal{C}$. Then, object accuracy for an image $x$, generated by a sentence containing objects $A$ and $B$ is:*

$$\text{OA}(x, A, B) = \mathbb{1}_{h(x)}(\exists A \cap \exists B). \tag{1}$$

Note that, the oracle function $h$ here could be either a pluggable learned model or a human detecting the presence of objects mentioned in the sentence. In our experiments, we show results for both cases and a correlational analyses between the two. Object accuracy is agnostic to the relationship $R$, whose presence is instead captured in the VISOR metric.

**Definition 2 (VISOR)** *Let $R_{gen}$ be the generated spatial relationship, while $R$ is the ground-truth relationship mentioned in text. Then, for each image $x$,*

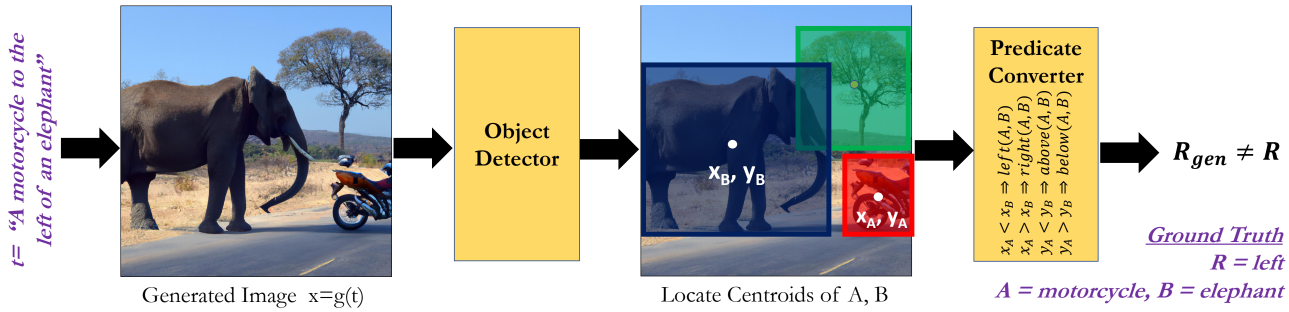



Figure 3: For text $t$ and corresponding generated image $x = g(t)$, object centroids are located and converted into predicates indicating the spatial relationship between them. These predicates are compared with the ground truth relationship $R$ to obtain the VISOR score.

$$\text{VISOR}(x, A, B, R) = \begin{cases} 1, \text{ if } (R_{gen} = R) \cap \exists A \cap \exists B \\ 0, \text{ otherwise.} \end{cases} \quad (2)$$

A useful feature of T2I models for artists and designers is the ability to generate multiple images for each input text prompt. This allows the creators to pick an appropriate image from $N$ generated images. We define $\text{VISOR}_\text{n}$ to reflect how good T2I models are at generating at least $n$ spatially correct images given a text input that mentions a spatial relationship. From a usability perspective (where creators have the option to pick from the output image set), $\text{VISOR}_\text{n}$ is useful for measuring if it is possible to find at least $n$ images that satisfy the prompt.

**Definition 3 ($\text{VISOR}_\text{n}$)** $\text{VISOR}_n$ *is the probability of generating images such that for every text prompt t, at least* $n$ *out of* $N$ *images have* VISOR=*1:*

$$\text{VISOR}_n(x, A, B, R) = \begin{cases} 1, & \sum_{i=1}^{N} \text{VISOR}(x_i, A, B, R) \geq n \\ 0, & \textit{otherwise}. \end{cases} \quad (3)$$

The relationship between VISOR and $\text{VISOR}_\text{n}$ is given below. The proof is presented in the supplementary materials.

$$\text{VISOR} = \frac{1}{N} \sum_{n=0}^{N} n(\text{VISOR}_n - \text{VISOR}_{n+1}). \quad (4)$$

In our study we use $N = 4$ images per text prompt and, therefore, report $\text{VISOR}_1$, $\text{VISOR}_2$, $\text{VISOR}_3$, and $\text{VISOR}_4$. Fig. 2 shows an example computation of all VISOR metrics.

Note that VISOR $= 1$ only if both objects are generated in the image, i.e. OA $= 1$. However, as we will see in Sec. 5, T2I models fail to generate multiple objects in a large subset of images. As such, it is important to disentangle the two abilities of the models to (1) generate multiple objects and (2) to generate them according to the spatial relationships described in the text of the prompt. For this purpose, we define conditional VISOR:

**Definition 4 (Conditional VISOR)** *is defined as the conditional probability of correct spatial relationships being generated, given that both objects were generated correctly.*

$$\text{VISOR}_{cond} = P(R_{gen}{=}R | \exists A \cap \exists B) \quad (5)$$

**Implementation.** The VISOR computation process is summarized in Fig. 3. Given any text prompt $t$ and a T2I model $g$, we first generate images $x = g(t)$, and use an object detector to localize objects in $x$. Object accuracy OA is computed using Eq. (1). We obtain centroid coordinates of objects A and B from the the bounding boxes of the detected objects. Based on the centroids, we deduce the spatial relationship $R_{gen}$ between them using the rules shown in the "Predicate Converter" box in Fig. 3. Finally, the generated relationship is compared with the ground-truth relationship $R$, and VISOR scores are computed using Eqs. (2), (3) and (5).

We use OWL-ViT [31], a state of the art open-vocabulary object detector, with a CLIP backbone and ViT-B/32 transformer architecture and confidence threshold 0.1. The supplementary material also contains results using DETR-ResNet-50 [5] trained on MS-COCO. The results using both object detectors are similar and lead to an identical ranking of models in our benchmark. However, the open-vocabulary functionality of OWL-ViT ensures that VISOR is widely applicable to other datasets, categories, and vocabularies. This removes dependence on specific datasets, making VISOR widely applicable for any freeform text input.

## 5. Experiments

**Baselines.** We study state-of-the-art T2I models as baselines: GLIDE [33], DALLE-mini [10], CogView2 [11], DALLE-v2 [36], and Stable-Diffusion (SD and SD 2.1.) [39], and two versions of Composable Diffusion Models [27] (GLIDE + CDM and SD + CDM). We generate $N$=4 images for each text prompt from our $\text{SR}_\text{2D}$ dataset, to obtain 126,720 images per model and compare performance in terms of OA, VISOR, $\text{VISOR}_\text{cond}$, and $\text{VISOR}_\text{1/2/3/4}$.

### 5.1. Ineffectiveness of Existing Metrics

T2I models have been primarily compared in terms of photorealism (purely visual) and human judgment about image quality (subjective). We quantify whether existing automated multimodal metrics are useful for evaluating spatial relationships generated by T2I models. We consider CLIP-

| Model | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE | CIDER | SPICE | CLIPScore | VISOR |
|---|---|---|---|---|---|---|---|---|---|---|
| GLIDE | 0.29 / 0 | 0.14 / 1.9e-6 | 0.05 / 2.9e-6 | 0.02 / 7.4e-7 | 0.13 / -9.9e-6 | 0.35 / -2.6e-6 | 0.18 / 1.1e-6 | 0.11 / -8.1e-6 | 0.70 / 1.3e-3 | 0.02 / 0.03 |
| GLIDE + CDM | 0.31 / 0 | 0.15 / 2.9e-5 | 0.06 / 7.6e-6 | 0.02 / 1.7e-6 | 0.15 / 6.4e-5 | 0.36 / 1.7e-6 | 0.22 / 3.1e-5 | 0.14 / -7.7e-5 | 0.75 / -8.8e-6 | 0.06 / 0.07 |
| DALLE-mini | 0.34 / 0 | 0.19 / -4.4e-5 | 0.09 / -1.6e-5 | 0.04 / -4.5e-6 | 0.19 / 2.1e-6 | 0.41 / -7.3e-6 | 0.34 / -6.0e-5 | 0.20 / 3.3e-5 | 0.80 / 1.5e-3 | 0.16 / 0.22 |
| CogView-2 | 0.30 / 0 | 0.16 / 4.4e-6 | 0.07 / 1.3e-6 | 0.03 / 5.6e-6 | 0.16 / -6.6e-6 | 0.36 / -2.9e-6 | 0.25 / 8.7e-6 | 0.15 / 7.0e-5 | 0.72 / 1.5e-5 | 0.12 / 0.13 |
| DALLE-v2 | 0.36 / 0 | 0.21 / -1.9e-5 | 0.11 / -4.8e-6 | 0.04 / -1.5e-6 | 0.21 / 1.7e-4 | 0.44 / 8.8e-6 | 0.40 / -2.8e-5 | 0.22 / -4.1e-5 | 0.84 / 1.8e-3 | 0.38 / 0.55 |
| SD | 0.33 / 0 | 0.18 / 3.3e-6 | 0.08 / 9.7e-7 | 0.03 / 2.8e-7 | 0.19 / 1.0e-5 | 0.40 / -2.6e-6 | 0.31 / 4.3e-6 | 0.19 / 7.4e-5 | 0.79 / 1.5e-3 | 0.19 / 0.23 |
| SD + CDM | 0.32 / 0 | 0.17 / 1.1e-5 | 0.07 / 5.1e-6 | 0.03 / 1.3e-6 | 0.17 / 1.6e-4 | 0.38 / 4.4e-6 | 0.28 / 1.2e-5 | 0.18 / -4.5e-5 | 0.77 / 3.6e-4 | 0.15 / 0.17 |
| SD 2.1 | 0.35 / 0 | 0.20 / -1.3e-5 | 0.09 / 4.2e-6 | 0.038 / -1.3e-6 | 0.20 / 7.1e-5 | 0.42 / 5.4e-6 | 0.35 / -1.8e-5 | 0.20 / 3.5e-5 | 0.82 / 1.0e-3 | 0.30 / 0.37 |

Table 2: $s/\Delta_s$ scores for T2I metrics shown in the 0 to 1 range. All previous metrics have low $\Delta_s$ (magenta) whereas VISOR has high $\Delta_s$ (green), showing they are ineffective in quantifying and benchmarking spatial understanding.

| Model | OA (%) | VISOR (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | uncond | cond | 1 | 2 | 3 | 4 |
| GLIDE [33] | 3.36 | 1.98 | 59.06 | 6.72 | 1.02 | 0.17 | 0.03 |
| GLIDE + CDM [27] | 10.17 | 6.43 | 63.21 | 20.07 | 4.69 | 0.83 | 0.11 |
| DALLE-mini [10] | 27.10 | 16.17 | 59.67 | 38.31 | 17.50 | 6.89 | 1.96 |
| CogView2 [11] | 18.47 | 12.17 | **65.89** | 33.47 | 11.43 | 3.22 | 0.57 |
| DALLE-v2 [36] | **63.93** | **37.89** | 59.27 | **73.59** | **47.23** | **23.26** | **7.49** |
| SD [39] | 29.86 | 18.81 | 62.98 | 46.60 | 20.11 | 6.89 | 1.63 |
| SD + CDM [27] | 23.27 | 14.99 | 64.41 | 39.44 | 14.56 | 4.84 | 1.12 |
| SD 2.1 | 47.83 | 30.25 | 63.24 | 64.42 | 35.74 | 16.13 | 4.70 |

Table 3: Comparison of the performance of all models in terms of object accuracy (OA) and each version of VISOR.

Score [16] (cosine similarity between image and text embeddings) and image captioning-based evaluation (BLEU [35], METEOR [4], ROUGE [24], CIDER [42], SPICE [2]) which are used by generating a caption $c$ for the synthesized image $x = g(t)$ and computing the captioning score with respect to the reference input text $t$. Note that purely visual metrics (FID and Inception Score [17, 41]) ignore the text, while semantic object accuracy [18] ignores all words except nouns, making them incapable of scoring spatial relationships.

Let $s^t$ be the score for $(x, t)$ where $x$ is the generated image and $t$ is the input text. Let $t_{flip}$ be the transformed version of $t$ obtained by inverting/flipping the spatial relationship in $t$ (for example, left→right). Let $s^t_{flip}$ be the score for $(x, t_{flip})$. For each metric, we define $\Delta_s$ as the average difference between $s^t$ and $s^t_{flip}$ over the entire $\mathrm{SR}_{\mathrm{2D}}$ dataset:

$$\Delta_s = \mathbb{E}_t[s^t - s^t_{flip}], \tag{6}$$

Thus, $\Delta_s$ captures the ability of metric $s$ to understand spatial relationships. Table 2 shows $s$ and $\Delta_s$ values for each previous metric and VISOR for each model. It can be seen that, for all previous metrics, $\Delta_s$ is negligible and close to zero, which implies that they return similar scores even if the text is flipped. For some cases, the difference is negative, implying that the score for the image and the flipped caption is higher. On the other hand, the $\Delta$ values for VISOR are high implying that VISOR assigns significantly lower scores for the flipped samples. These results establish the need for a new evaluation metric since none of the existing metrics are able to quantify spatial relationships reliably, and show the efficacy of VISOR for this purpose.

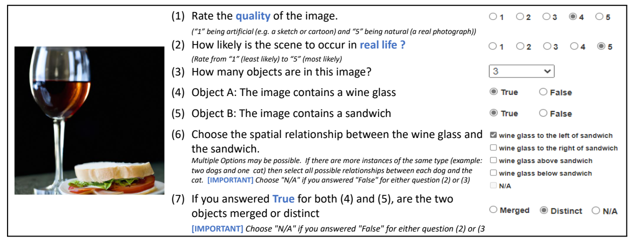



Figure 4: The human study interface with an image on the left and seven multiple choice questions about it.

## 5.2. Benchmarking Results

Table 3 shows the results of benchmarking on our $\mathrm{SR}_{\mathrm{2D}}$ dataset. We first note that the object accuracy of all models except DALLE-v2 is lower than 30%. While DALLE-v2 (63.93%) significantly outperforms other models, it still shows a large number of failures in generating both objects that are mentioned in the prompt. For the unconditional metrics VISOR and $\mathrm{VISOR}_{1/2/3/4}$, DALLE-v2 is the best performing model. However, in terms of $\mathrm{VISOR}_{\mathrm{cond}}$, CogView2 has the highest performance. This implies that, although CogView2 is better than other models on those examples where both objects are generated, the large failures of CogView2 in OA result in a lower unconditional VISOR score. $\mathrm{VISOR}_4$ is extremely low for all models including DALLE-v2 ($8.54\%$), revealing a large gap in performance.

## 5.3. Human Study

**Methodology.** We conducted a human evaluation study to understand the alignment of our metrics with human judgment, and to quantify the gap between object detector performance and human assessments of object presence. For the human study, we used four models: CogView2, DALLE-v2, Stable Diffusion (SD), and SD + CDM. Annotators were shown (via Amazon Mechanical Turk) an image generated by one of the four models, and were asked seven questions about it, as shown in Fig. 4. The questions assessed human evaluation of image quality and scene realism (*scene likelihood*) on a Likert scale (1 through 5), the number of objects, answering True or False for presence of objects, selecting valid spatial relationships, and responding if two objects

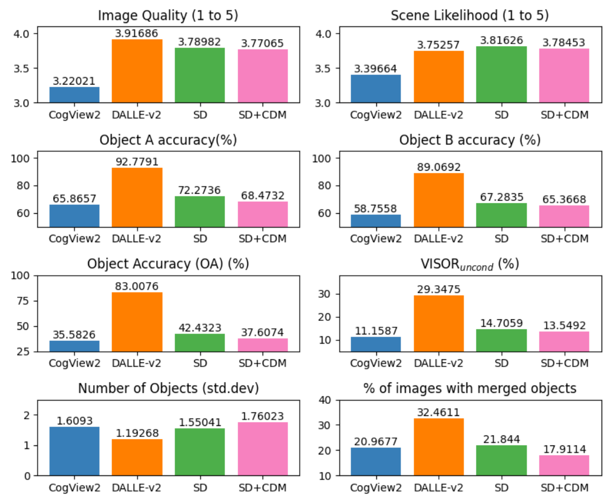



Figure 5: Summary of responses to each question in the human study, compared across all four models.

| Response | CogView2 | DALLE-v2 | SD | SD + CDM |
|---|---|---|---|---|
| Image Quality | 65.47 / 52.93 | 75.02 / 62.33 | 69.86 / 55.31 | 72.59 / 57.99 |
| Scene Likelihood | 64.40 / 50.78 | 72.13 / 59.62 | 69.47 / 52.35 | 67.19 / 53.99 |
| Num. Objects | 79.63 / 50.03 | 87.09 / 46.39 | 81.41 / 46.06 | 80.28 / 45.74 |
| Object A | 100.0 / 33.00 | 99.64 / 8.02 | 100.0 / 18.56 | 100.0 / 20.04 |
| Object B | 100.0 / 32.75 | 100.0 / 13.39 | 100.0 / 22.44 | 100.0 / 25.51 |
| Spatial Relation | 100.0 / 23.33 | 100.0 / 47.90 | 100.0 / 30.79 | 100.0 / 25.00 |
| Merged/Distinct | 100.0 / 43.02 | 99.64 / 58.85 | 100.0 / 39.95 | 100.0 / 38.60 |

Table 4: Majority / Unanimous inter-worker agreement (%) for each question in our human study.

| Metric | CogView2 | DALLE-v2 | SD | SD-CDM |
|---|---|---|---|---|
| OA | 73.07 | 73.87 | 79.25 | 80.21 |
| VISOR$_{\text{uncond}}$ | 88.48 | 77.41 | 88.43 | 88.80 |
| VISOR$_{\text{cond}}$ | 75.02 | 75.62 | 76.95 | 74.69 |

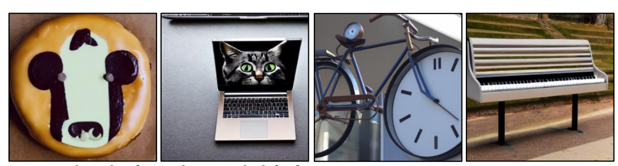
Table 5: Agreement(%) of human responses with automated metrics

were merged in the image. We used a sample size of 1000 images per model and 3 workers per sample.

**Results.** Fig. 5 shows a summary of responses for each question in the human study. While DALLE-v2 received the highest image quality rating, SD and SD+CDM received higher scene likelihood rating. Interestingly, DALLE-v2 also had the largest number of images with merged objects (32.46%); several cases of this phenomenon are shown in Fig. 7. Inter-annotator agreement was high for all questions in terms of majority (agreement between at least 2 out of 3 workers) and unanimous agreement (agreement between all 3 out of 3 workers) as reported in Table 4.

**Alignment of VISOR with Human Responses.** We observe that the ranking of models in terms of both object accuracy (OA) and VISOR is identical for the human study and for the automated VISOR scores in Table 3, i.e. *DALLE-v2 > SD > SD-CDM > CogView2*. Table 5 shows the percentage of samples for which responses from humans matched our automated evaluation using object detectors.


Figure 6: Illustrative examples of text prompts from our SR$_{2D}$ dataset and corresponding images generated by each T2I model.



a cow to the right of a donut | a laptop to the left of a cat | a clock above a bicycle | a bench below a keyboard

Figure 7: Illustrative examples where the two objects from the text input appear to be merged. From left to right: *a, b, c, d*.

## 6. Analysis

**Qualitative Results.** Fig. 6 shows examples of images generated by all baselines for each prompt, with more visualizations in the appendix. Although the photorealism of recent models, such as DALLE-v2, SD, and SD+CDM, is much higher, all models are equally poor at generating accurate spatial relationships.

**Merged Objects.** Fig. 7 shows examples of a few common types of merging between objects that we observed, espe-

| Model | VISOR$_{cond}$ (%) | | | | Object Accuracy (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | left | right | above | below | left | right | above | below |
| GLIDE | 57.78 | **61.71** | 60.32 | 56.24 | 3.10 | 3.46 | **3.49** | 3.39 |
| GLIDE + CDM | 65.37 | **65.46** | 59.40 | 59.84 | **12.78** | 12.46 | 7.75 | 7.68 |
| DALLE-mini | 57.89 | 60.16 | **63.75** | 56.14 | 22.29 | 21.74 | **33.62** | 30.74 |
| CogView2 | **68.50** | 68.03 | 63.72 | 62.51 | **20.34** | 19.30 | 17.71 | 16.54 |
| DALLE-v2 | 56.47 | 56.51 | 60.99 | **63.24** | 64.30 | 64.32 | **65.66** | 61.45 |
| SD | **64.44** | 62.73 | 61.96 | 62.94 | 29.00 | 29.89 | **32.77** | 27.8 |
| SD + CDM | **69.05** | 66.52 | 62.51 | 59.94 | 23.66 | 21.17 | 23.66 | **24.61** |

Table 6: Comparison of Visor and OA split by relationship type

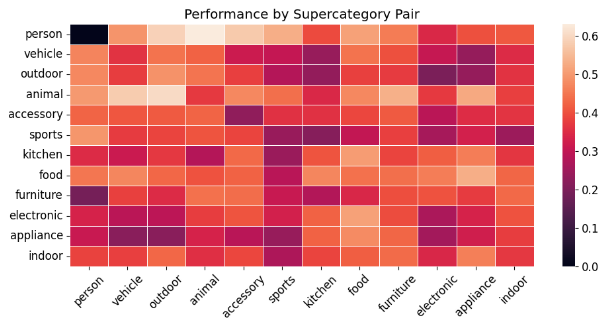



Figure 8: VISOR scores for each supercategory pair.

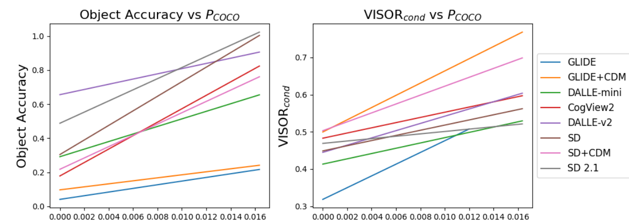



Figure 9: Correlation of our metrics with $P_{COCO}$, the object co-occurrence probability in MS-COCO.

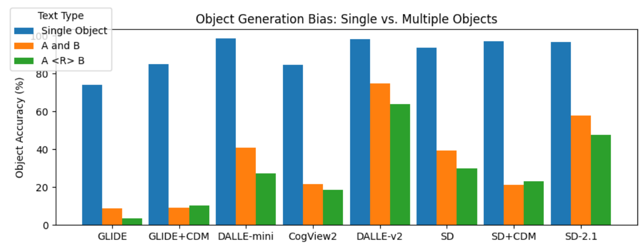



Figure 10: Comparison of object accuracy for text with single and multiple objects reveals a bias towards single objects.

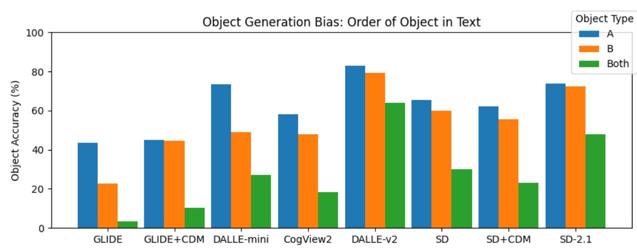



Figure 11: Comparison of object accuracy for object A and B reveals a bias towards A, the first object appearing in the prompt).

cially with DALLE-v2. Common patterns observed include animals being rendered as patterns on inanimate objects (a, b) and both objects retaining their typical shape but getting merged (c, d). As our human study in Fig. 5 shows, a large proportion (more than 20%) of images have merged objects – this poses a significant challenge for generating distinct objects and their relationships using T2I models.

**Performance per relationship** is shown in Table 6. Interestingly, five of the seven models have the best VISOR$_{cond}$ scores for horizontal relationships (left or right). However, five of the seven models have the best object accuracy for vertical relationships (above or below).

**Performance per Supercategory.** The 80 object categories in SR$_{2D}$ belong to 11 MS-COCO "supercategories". We investigate VISOR scores for each supercategory pair and report the results for the best model (DALLE-v2) in Fig. 8 (results for other models are in the appendix). VISOR scores for commonly co-occurring supercategories such as *"animal, outdoor"* are highest whereas unlikely combinations of indoor-outdoor objects such as *"vehicle, appliance"* and *"electronic, outdoor"* have low performance.

**Correlation between VISOR and Object Co-occurrence.** The object categories in our dataset span a wide range of commonly occurring objects from MS-COCO such as wild animals, vehicles, appliances, and humans, found in varying contexts, including combinations that do not appear together often in real life. For instance, an elephant is unlikely to be found indoors near a microwave oven. To understand how object co-occurrence affects VISOR, we first obtain $P_{\text{COCO}}(A, B)$, the probability of co-occurrence for each object pair (A, B) as a proxy for real-world object co-occurrence. Then, we plot the correlation of VISOR and object accuracy for pair (A, B) with its $P_{\text{COCO}}(A, B)$. As Fig. 9 shows, the correlation is positive for all models, for both OA and VISOR$_{cond}$, implying that the quality of outputs is likely to be better for commonly co-occurring objects, clearly establishing a bias towards real-world likelihood. This correlation shows the difficulty in generating unlikely relationships such as *"an elephant to the left of a microwave"* even though such unlikely combinations may be desired by creators, pursuing artistic compositions.

**Object Generation Bias.** We compare object accuracy with three types of inputs to generate images: (1) single objects text such as *"an elephant"*, (2) multiple object conjunction such as *"an elephant and a cat"*, and (3) relational texts such as *"an elephant to the right of a cat"*. Fig. 10 shows that, for all models, OA is significantly higher for single objects; while composition using conjunction is challenging, systems

| **Model** | left | right | above | below | Average |
|---|---|---|---|---|---|
| GLIDE | <u>45.90</u> | 58.93 | 63.16 | 52.63 | 55.16 |
| GLIDE + CDM | 61.99 | 59.15 | 54.79 | 56.15 | 58.02 |
| DALLE-mini | 54.75 | 52.28 | 54.64 | 55.77 | 54.36 |
| CogView2 | **67.32** | 65.38 | **65.67** | **66.95** | **66.33** |
| DALLE-v2 | 48.81 | <u>48.10</u> | <u>48.72</u> | <u>48.15</u> | <u>48.45</u> |
| SD | 58.71 | 61.36 | 55.36 | 55.39 | 57.71 |
| SD + CDM | 64.69 | **65.71** | 61.35 | 57.71 | 62.37 |
| SD 2.1 | 53.96 | 55.50 | 54.73 | 54.38 | 54.64 |

Table 7: Consistency (%) of generated spatial relationships for equivalent inputs. **Bold**: highest, <u>Underline</u>: lowest consistency.

| **Prompt Type** | **OA (%)** | **VISOR (%)** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **uncond** | **cond** | **1** | **2** | **3** | **4** |
| Phrases | 29.86 | 18.81 | 62.98 | 46.60 | 20.11 | 6.89 | 1.63 |
| Sentences | 32.48 | 20.67 | 63.64 | 48.54 | 22.94 | 8.92 | 2.25 |
| Split Sentences | 24.98 | 16.44 | 65.82 | 41.91 | 16.29 | 5.66 | 1.91 |

Table 8: Effect of prompt variations on OA and VISOR scores. All three versions use the same Stable Diffusion (SD) model .

perform better with this generation than spatial composition.

**Text-Order Bias.** In Fig. 11, we show that for all models, OA for the first mentioned object (A) in the text is significantly higher than OA for the second object (B); generating both objects together is most challenging.

**Consistency between equivalent phrases.** Ideally, given two equivalent inputs such as *"a cat above a dog"* and *"a dog below a cat"*, the model should generate images with the same spatial relationship. To evaluate this consistency, we consider cases in which both objects are detected and report the consistency for each relationship type in Table 7. Surprisingly, the best model *DALLE-v2 is the least consistent*, while CogView2 is the most consistent model. This result shows that merely rephrasing the input can have a large influence on the spatial correctness of the output.

**Effect of Attributes on Spatial Understanding.** We conduct a case study with Stable Diffusion (SD) to seek an understanding of the impact of sentence complexity on a model's VISOR performance. We increase the complexity of text prompts by randomly assigning two attributes (size `Z` and color `C`) to the object category, via templates of the form `[`$Z_A$`][`$C_A$`]<A> <R> [`$Z_B$`][`$C_B$`]<B>`. We focus on 11 object categories representative of each supercategory in COCO, 8 colors, and 4 sizes. As shown in Fig. 12, compared to generation without attributes, there is a drop in performance in 13 out of 15 types of attribute combinations. Addition of the color attribute (C) leads to a large drop in performance. Adding size descriptors (Z) may improve performance. While concurrent work [12] has reported difficulty in attribute binding, our analysis suggests that attributes may negatively influence spatial compositionality.

**Effect of Rephrased Text Prompts.** We compare variations

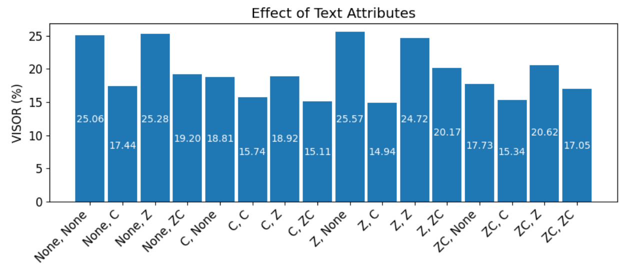



Figure 12: Comparing VISOR performance with different combinations of attributes. *"Z, ZC"* indicates a prompt describing object A with a size attribute and object B with both size and color.

of the prompt: (1) *phrases*: the default version of $SR_{2D}$ used in Tab. 3 (*e.g.* "a cat to the left of a dog"); (2) *sentences* (*e.g.* "There is a cat to the left of a dog"), and (3) *split sentences* (*e.g.* "There is a cat to the left. There is a dog to the right"). Compared to phrases, Tab. 8 shows higher OA and $\text{VISOR}_{\text{cond}}$ for *sentences*; lower OA and higher $\text{VISOR}_{\text{cond}}$ for *split sentences*. Prompt engineering for grounded generation such as spatial aspects is a promising future direction.

## 7. Discussion and Conclusion

We studied the spatial capabilities of text-to-image generators by introducing spatial relationship metrics (VISOR measures), building a dataset ($SR_{2D}$), and developing an automated evaluation pipeline. Our experiments reveal that existing T2I models have poor spatial interpretation and rendering abilities, as characterized by their low VISOR scores, making them unreliable for uses that depend on the correctness in generated images of spatial relations specified in prompts. Our analysis also reveals several biases and artifacts of T2I models, such as proclivity for generating single objects (especially the first mentioned object), correlation of spatial correctness with likelihood of object co-occurrence, sensitivity to equivalent phrasings of spatial relations, and negative influences of the inclusion in prompts of several commonly used modifiers. We hope that the metrics, methods, and dataset will help to stimulate a stream of research on the spatial rendering capabilities of generative models, leading to enhancements of these capabilities over time. For uses of today's technologies, we hope our findings can provide creators with guidance for prompt engineering. We note that the $SR_{2D}$ data generation pipeline can be extended to study spatial relationships of more than two objects, including three-dimensional and complex relations such as *inside, outside, contains, behind, in front, covers, touching*, as well as semantic and action-based relationships. We hope the VISOR metric will serve as a complement to prior metrics for evaluating photorealism and image-text similarity.

## References


[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. IEEE Computer Society, 2018. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 5

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 1

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 5

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4, 12

[6] 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, 2023. 12, 13

[7] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019. 2

[8] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*, 2022. 3

[9] Colin Conwell and Tomer Ullman. Testing relational understanding in text-guided image generation. *arXiv preprint arXiv:2208.00005*, 2022. 2

[10] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 7 2021. 2, 4, 5, 13

[11] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2, 4, 5, 12, 13

[12] 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. 8, 12, 13

[13] Vatsal Goel, Mohit Chandak, Ashish Anand, and Prithwijit Guha. Iq-vqa: intelligent visual question answering. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 357–370. Springer, 2021. 2

[14] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 379–396. Springer, 2020. 2

[15] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 2

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 1, 5

[17] 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. *Advances in neural information processing systems*, 30, 2017. 1, 5

[18] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1, 5

[19] Seunghoon Hong, Xinchen Yan, Thomas E. Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2713–2723, 2018. 1

[20] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

[21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society, 2017. 2

[22] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2785, 2021. 2

[23] Evelina Leivada, Elliot Murphy, and Gary Marcus. Dall-e 2 fails to reliably capture common syntactic processes. *arXiv preprint arXiv:2210.12889*, 2022. 3

[24] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 3

[26] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 2

[27] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. *European Conference on Computer Vision*, 2022. 2, 4, 5, 12, 13

[28] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4185–4194. Computer Vision Foundation / IEEE, 2019. 2

[29] Gary Marcus, Ernest Davis, and Scott Aaronson. A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*, 2022. 3

[30] Oier Mees, Alp Emek, Johan Vertens, and Wolfram Burgard. Learning object placements for relational instructions by hallucinating scene representations. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 94–100. IEEE, 2020. 1

[31] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. 4, 12

[32] Suraj Nair, Eric Mitchell, Kevin Chen, Silvio Savarese, Chelsea Finn, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*, pages 1303–1315. PMLR, 2022. 1

[33] 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 *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2, 4, 5, 13

[34] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 2

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 5

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2, 4, 5, 12, 13

[37] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5860–5865, Hong Kong, China, 2019. Association for Computational Linguistics. 2

[38] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 2

[39] 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*, pages 10684–10695, 2022. 2, 4, 5, 12, 13

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 3

[41] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1, 5

[42] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[43] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical Report CNS-TR-201, Caltech, 2010. 2

[44] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 1

[45] Keren Ye and Adriana Kovashka. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3181–3189, 2021. 2

[46] 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. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification. 2

[47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[48] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, 2017. Association for Computational Linguistics. 2

# Appendix

In this appendix we provide more details about the $\text{SR}_{\text{2D}}$ dataset, additional experimental results and analyses, updates to the benchmark, proof of 4, and qualitative examples.

## A. Additional Details about the $\text{SR}_{\text{2D}}$ Dataset

**List of COCO categories.** In the $\text{SR}_{\text{2D}}$ dataset we use 80 object categories from the MS-COCO dataset as the set of objects $\mathcal{C}$. The box below lists all of these categories.

```
person, bicycle, car, motorcycle, airplane,
bus, train, truck, boat, traffic light,
fire hydrant, stop sign, parking meter,
bench, bird, cat, dog, horse, sheep, cow,
elephant, bear, zebra, giraffe, backpack,
umbrella, handbag, tie, suitcase, frisbee,
skis, snowboard, sports ball, kite,
baseball bat, baseball glove, skateboard,
surfboard, tennis racket, bottle, wine
glass, cup, fork, knife, spoon, bowl,
banana, apple, sandwich, orange, broccoli,
carrot, hot dog, pizza, donut, cake, chair,
couch, potted plant, bed, dining table,
toilet, tv, laptop, mouse, remote, keyboard,
cell phone, microwave, oven, toaster, sink,
refrigerator, book, clock, vase, scissors,
teddy bear, hair drier, toothbrush
```

**List of COCO supercategories.** In Figure 8, we presented results for each supercategory. The box below lists these eleven supercategories.

```
person, vehicle, outdoor, animal, accessory,
sports, kitchen, food, furniture,
electronic, appliance, indoor
```

**Experimental Setup for "Effect of Attributes on Spatial Understanding".** In Figure 12, we compared VISOR scores for text prompts with and without size and color attributes. We used one object category for 11 supercategories for this analysis. Note that we ignore the `person` category since colors are not typically used as attributes for people (for example "purple person") and to avoid any potentially racist stereotypes associated with skin color to percolate into our generated images. Table 9 shows the object categories that we used:

## B. Proof of Equation 4

Equation 4 states the following relationship between VISOR and $\text{VISOR}_{\text{n}}$.

$$\text{VISOR} = \frac{1}{N}\sum_{n=0}^{N} n(\text{VISOR}_n - \text{VISOR}_{n+1}). \qquad (7)$$

| **Supercategory** | vehicle | outdoor | animal | accessory | sports | kitchen | food | furniture | electronic | appliance | indoor |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Category** | car | bench | dog | suitcase | sports ball | cup | cake | chair | laptop | microwave | book |

Table 9: Object categories that we used for each supercategory for analyzing the effrt of attributes on spatial understansing.

**Proof.** First, we restate the definitions of VISOR and VISOR$_n$ below.

$$\text{VISOR}(x, A, B, R) = \begin{cases} 1, & \text{if } R_{gen} = R \cap A \cap B \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

$$\text{VISOR}_n(x, A, B, R) = \begin{cases} 1, & \text{if } \sum_{i=1}^{N} \text{VISOR}(x_i, A, B, R) \geq n \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Let $T$ be the total number of text prompts used for evaluating VISOR of a text-to-image model. For a model that generates $N$ images per prompt, we have $NT$ total generated images. Let $V$ be the number of images for which VISOR $= 1$, i.e. images for which $R = R_{gen} \cap A \cap B$. From Eq. (8), it is clear that

$$\text{VISOR} = \frac{V}{NT}. \tag{10}$$

Let $P_n$ be the number of prompts for which *at least* $n$ generated images were spatially correct. From Eq. (9) we can say:

$$\text{VISOR}_n = \frac{P_n}{T}$$

$\Rightarrow P_n - P_{n+1}$ is the number of prompts for which *exactly* $n$ images are correct.
$\Rightarrow \sum_{i=0}^{N} n(P_n - P_{n+1})$ is the number of generated images which are spatially correct.

$$\Rightarrow V = \sum_{i=0}^{N} n(P_n - P_{n+1})$$

$$\Rightarrow \frac{V}{NT} = \frac{1}{NT} \sum_{i=0}^{N} n(P_n - P_{n+1})$$

$$\Rightarrow \text{VISOR} = \frac{1}{N} \sum_{i=0}^{N} n(\text{VISOR}_n - \text{VISOR}_{n+1})$$

*from Eqs. (9) and (10)*

## C. Additional Experiments

**Performance Per Supercategory.** Fig. 15 shows the performance per supercategory for all seven models.

**Benchmarking using COCO-finetuned Object Detectors.** In the main paper, we used the open-vocabulary object detector OWL-ViT [31] as the oracle to localize objects and identify their spatial relationships. In this section, we replicate these benchmarking experiments by using DETR-Resnet-50 [5], finetuned on the MS-COCO dataset. Results are shown in Tab. 10. The findings are similar using OWLViT. In terms of OA and VISOR (unconditional), DALLE-v2 [36] is the best performing model. However conditional VISOR for DALLE-v2 is low, but higher for CogView [11], SD [39], and SD+CDM [27]. This result shows that irrespective of the object detector, the relative performance comparison and rankings of models obtained via VISOR are consistent. While the oracle object detector does set an upper-bound for VISOR, it can be seamlessly swapped with any newer and more sophisticated object detectors that may be developed in the future.

| **Model** | **OA (%)** | **VISOR (%)** uncond | cond | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|
| GLIDE | 0.23 | 2.54 | 0.12 | 0.47 | 0.02 | 0.00 | 0.00 |
| GLIDE+CDM | 1.49 | 5.09 | 0.82 | 2.90 | 0.33 | 0.04 | 0.01 |
| DALLE-mini | 6.91 | 3.16 | 3.67 | 11.34 | 2.65 | 0.62 | 0.08 |
| CogView2 | 4.75 | 6.86 | 2.70 | 8.85 | 1.62 | 0.29 | 0.04 |
| DALLE-v2 | 14.80 | 3.92 | 7.98 | 22.84 | 6.95 | 1.80 | 0.32 |
| SD | 14.17 | 7.14 | 8.09 | 23.37 | 6.91 | 1.73 | 0.37 |
| SD+CDM | 11.06 | 9.34 | 6.56 | 19.49 | 5.13 | 1.40 | 0.22 |

Table 10: Benchmarking performance of all models in terms of object accuracy (OA) and each version of VISOR, with DETR-ResNet-50 (trained on MS-COCO) as the object detector

**Effect of Confidence Threshold.** We study how the confidence threshold of the object detector affects VISOR performance. In Fig. 13 we plot the VISOR scores for each model for four values of confidence threshold: 0.1, 0.2, 0.3, and 0.4. For a higher (stricter) threshold, naturally, the VISOR score is lower, since fewer objects will be detected than at a lower confidence value. However, the trend and relative performance of the models are identical irrespective of the confidence threshold, leading to consistency in the rankings of models in terms of VISOR score, validating the use of oracle object detectors for computing VISOR score. This also implies that in the future, more sophisticated object detection models that may be developed can be incorporated in the VISOR computation pipeline to replace older detectors.

**`UPDATE`: Comparison with new spatially-focused methods.** After the first version of this preprint was released, more variants of text-to-image generation models have been developed. Two of these: Structured Diffusion [12] and Attend-and-Excite [6] are focused on spatial aspects of the image and are therefore very relevant to the benchmark. The availability of our dataset and metric will aid in comparing

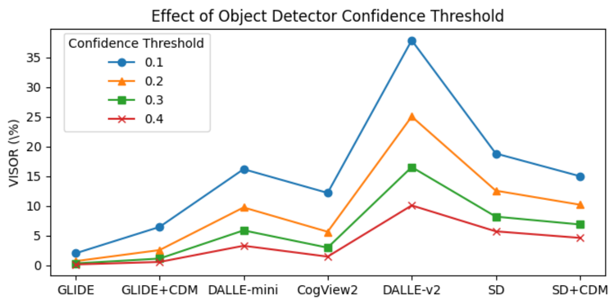



Figure 13: Effect of the confidence threshold of the oracle object detector. The relative ranking of models is consistent irrespective of the confidence threshold.

| Model | OA (%) | VISOR (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | uncond | cond | 1 | 2 | 3 | 4 |
| GLIDE [33] | 3.36 | 1.98 | 59.06 | 6.72 | 1.02 | 0.17 | 0.03 |
| GLIDE + CDM [27] | 10.17 | 6.43 | 63.21 | 20.07 | 4.69 | 0.83 | 0.11 |
| DALLE-mini [10] | 27.10 | 16.17 | 59.67 | 38.31 | 17.50 | 6.89 | 1.96 |
| CogView2 [11] | 18.47 | 12.17 | **65.89** | 33.47 | 11.43 | 3.22 | 0.57 |
| DALLE-v2 [36] | **63.93** | **37.89** | 59.27 | **73.59** | **47.23** | **23.26** | **7.49** |
| SD [39] | 29.86 | 18.81 | 62.98 | 46.60 | 20.11 | 6.89 | 1.63 |
| SD + CDM [27] | 23.27 | 14.99 | 64.41 | 39.44 | 14.56 | 4.84 | 1.12 |
| SD 2.1 | 47.83 | 30.25 | 63.24 | 64.42 | 35.74 | 16.13 | 4.70 |
| Structured Diffusion [12] | 28.65 | 17.87 | 62.36 | 44.70 | 18.73 | 6.57 | 1.46 |
| Attend-and-Excite [6] | 42.07 | 25.75 | 61.21 | 49.29 | 19.33 | 4.56 | 0.08 |

Table 11: Performance of new spatially-focused models on our benchmark.

these or future models in terms of spatial reasoning. Table 11 shows the comparitive results with the models in our previous study.

**Linguistic Variations with Large Language Models.** Our $\text{SR}_{2D}$ dataset offers a controlled evaluation without ambiguity or noise. However recent advances in generative language models have given us access to automated ways to rephrase our prompts. In a small experiment, we obtaining 3 variations from GPT3.5-Turbo for 500 $\text{SR}_{2D}$ prompts and generated images using Stable Diffusion 2.1. Examples of the GPT-rephrased text promps are shown in Tab. 12. We compare VISOR scores of GPT-rephrased vs. original prompts in Tab. 13 and find that with GPT-rephrased prompts, OA is higher but VISOR is lower – hinting that prompt engineering with LLMs might help object-level evaluation, but still may not enhance spatial understanding.

| Original | GPT-Variation |
|---|---|
| an apple below a skateboard | (1) Beneath the skateboard lies an apple.<br>(2) An apple rests underneath a skateboard. |
| a cat to the right of a toaster | (1) A cat positioned on the right side of a toaster.<br>(2) On the right side of a toaster, there is a cat. |

Table 12: GPT-rephrased versions of $\text{SR}_{2D}$ prompts.

| Dataset | OA (%) | VISOR (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | uncond | cond | 1 | 2 | 3 | 4 |
| Original | 45.83 | 30.47 | 66.49 | 65.69 | 35.62 | 16.34 | 4.25 |
| GPT-rephrased | 46.16 | 29.99 | 64.98 | 63.54 | 35.10 | 16.59 | 4.74 |

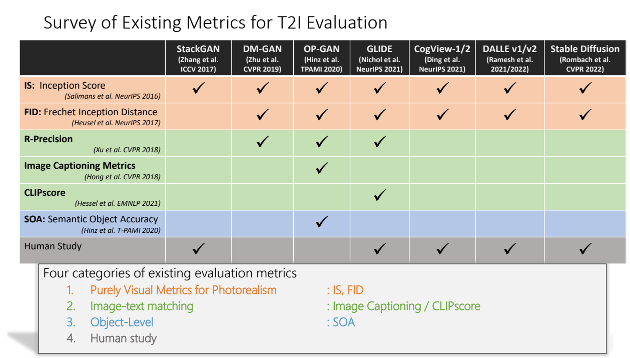
Table 13: Comparison between performance of SD2.1. on the original $\text{SR}_{2D}$ dataset and its GPT-rephrased versions.

## D. Survey of Prior Work on T2I Evaluation

Text-to-image synthesis is a relatively new area of research but has seen an explosion in interest and unprecedented improvements in the quality of generated outputs. In Fig. 14 we overview existing evaluation metrics for T2I and their used by seminal T2I models from 2017 to 2022. We categorize these metrics into four broad categories: purely visual metrics, image-text matching, object-level evaluation, and human studies.

## E. Visualization of Generated Images

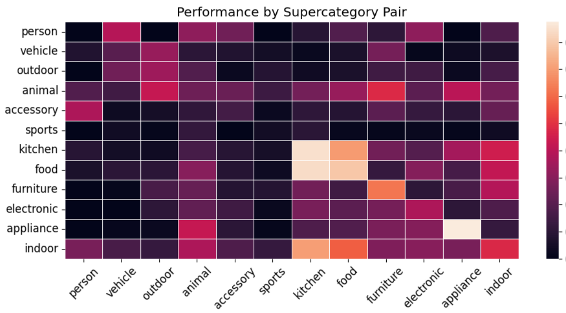
We would like readers to view more examples via our anonymous project page `visort2i.github.io`, or the webpage `viz.html` found in the accompanying `.zip` file. These webpages contain additional visualizations of images generated by our benchmark models for text prompts from the $\text{SR}_{2D}$ dataset. In this section we visualize some of these examples and show all $N = 4$ images generated by each model, in Figs. 16 to 23.


Survey of Existing Metrics for T2I Evaluation

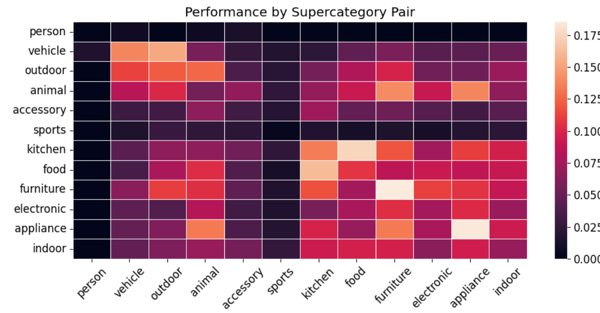
| | StackGAN (Zhang et al. ICCV 2017) | DM-GAN (Zhu et al. CVPR 2019) | OP-GAN (Hinz et al. TPAMI 2020) | GLIDE (Nichol et al. NeurIPS 2021) | CogView-1/2 (Ding et al. NeurIPS 2021) | DALLE v1/v2 (Ramesh et al. 2021/2022) | Stable Diffusion (Rombach et al. CVPR 2022) |
|---|---|---|---|---|---|---|---|
| **IS:** Inception Score *(Salimans et al. NeurIPS 2016)* | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **FID:** Frechet Inception Distance *(Heusel et al. NeurIPS 2017)* | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **R-Precision** *(Xu et al. CVPR 2018)* | | ✓ | ✓ | ✓ | | | |
| **Image Captioning Metrics** *(Hong et al. CVPR 2018)* | | | ✓ | | | | |
| **CLIPscore** *(Hessel et al. EMNLP 2021)* | | | | ✓ | | | |
| **SOA:** Semantic Object Accuracy *(Hinz et al. T-PAMI 2020)* | | | ✓ | | | | |
| Human Study | ✓ | | | ✓ | ✓ | ✓ | ✓ |

Four categories of existing evaluation metrics
1. Purely Visual Metrics for Photorealism : IS, FID
2. Image-text matching : Image Captioning / CLIPscore
3. Object-Level : SOA
4. Human study

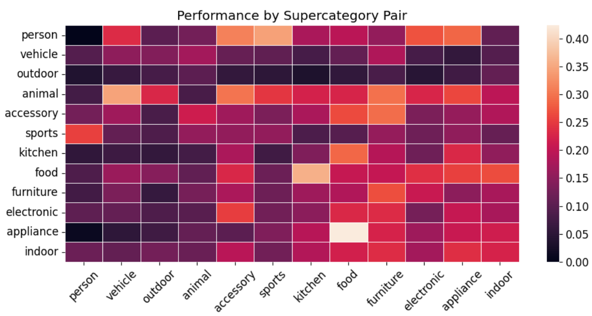
Figure 14: Overview of previous T2I evaluation metrics.



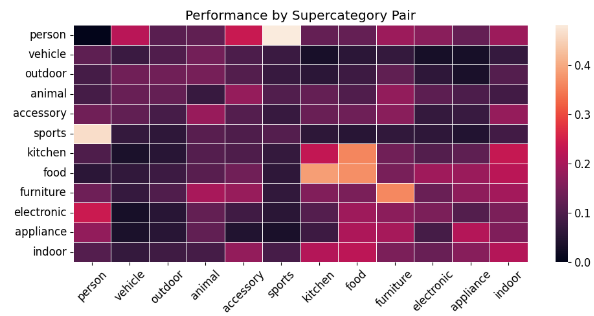

(a) GLIDE



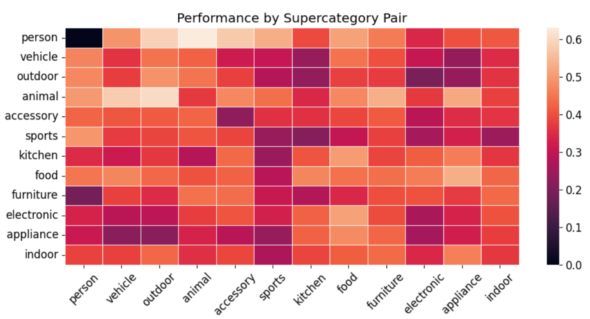

(b) GLIDE+CDM



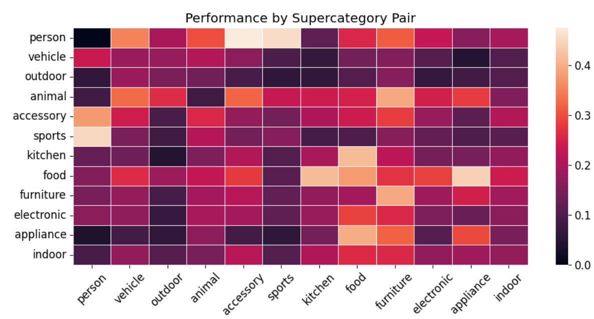

(c) DALLE-mini



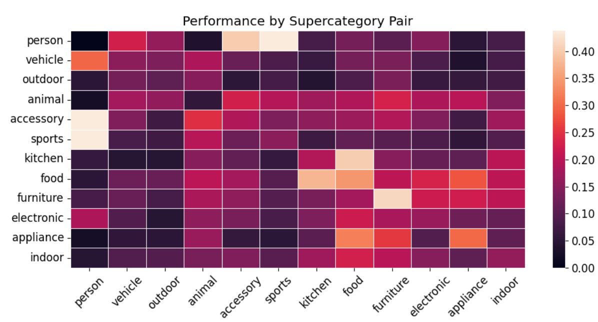

(d) CogView2




(e) DALLE-v2




(f) Stable Diffusion




(g) SD+CDM

Figure 15: VISOR scores of each model split by supercategory pairs.

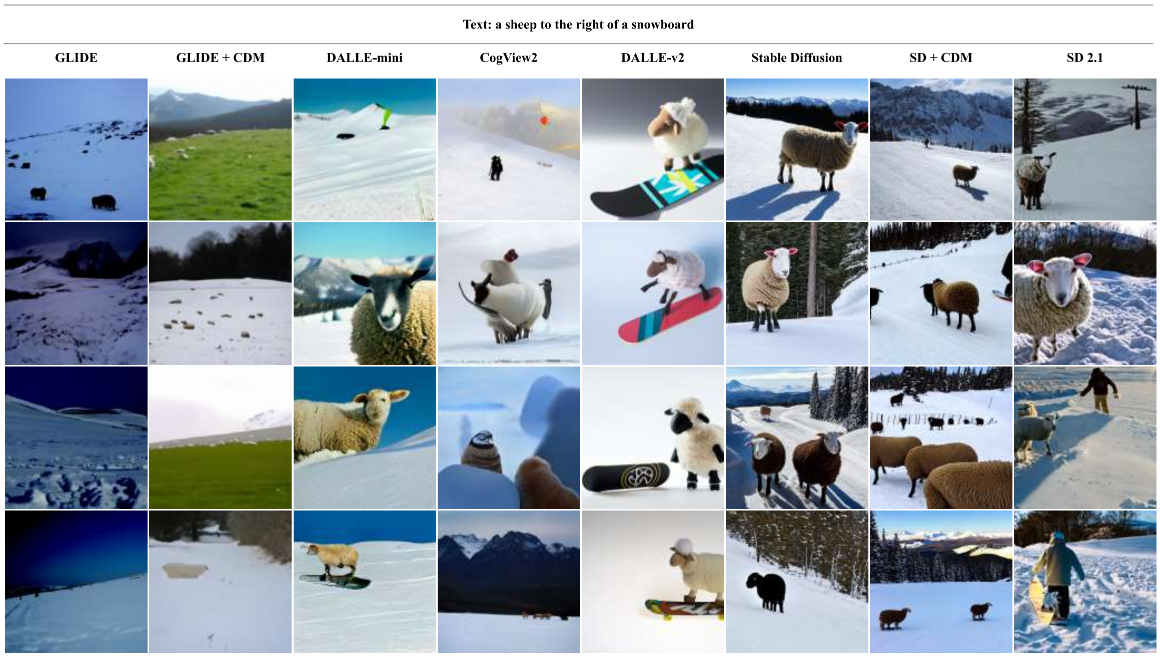



Figure 16: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.

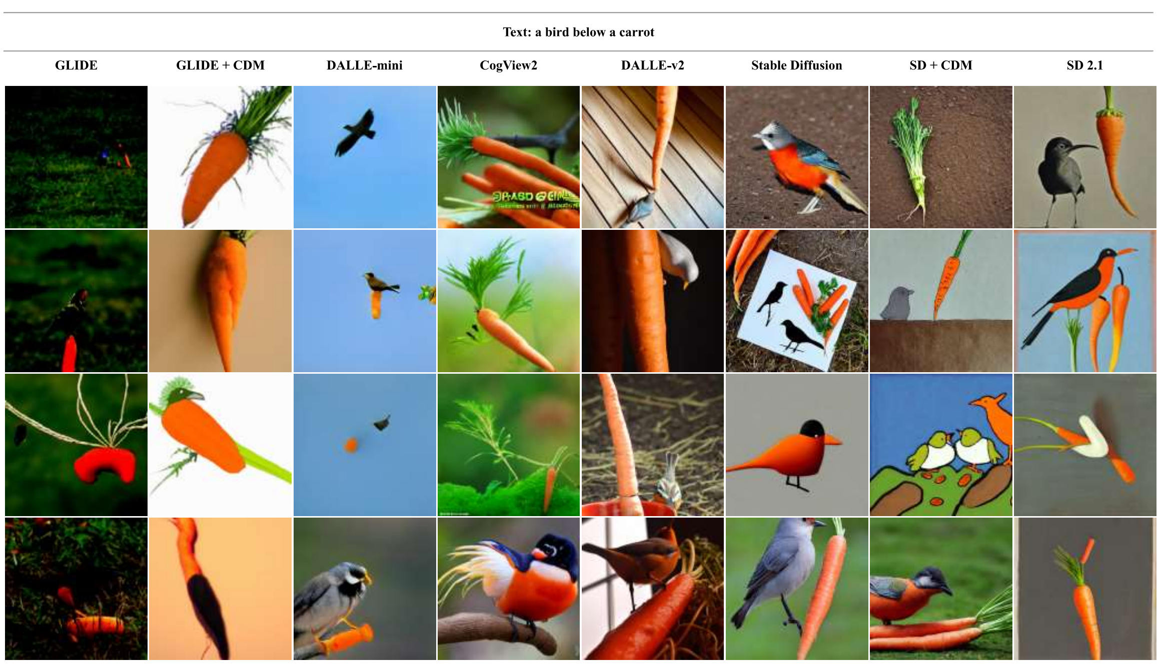



Figure 17: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.

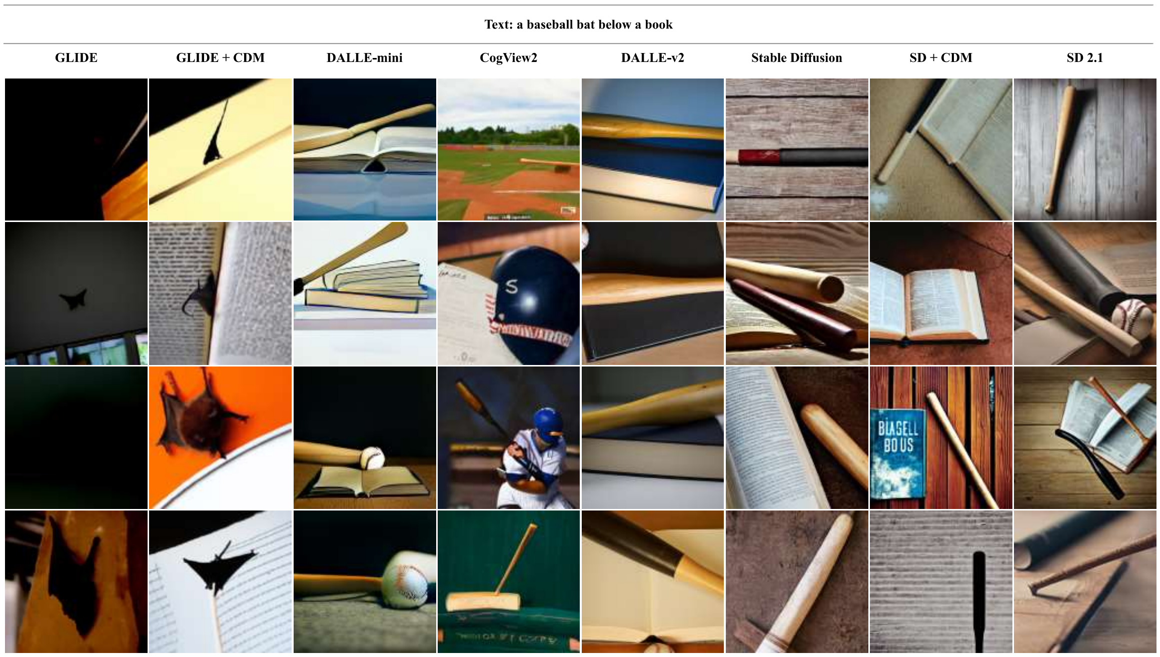



Figure 18: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.

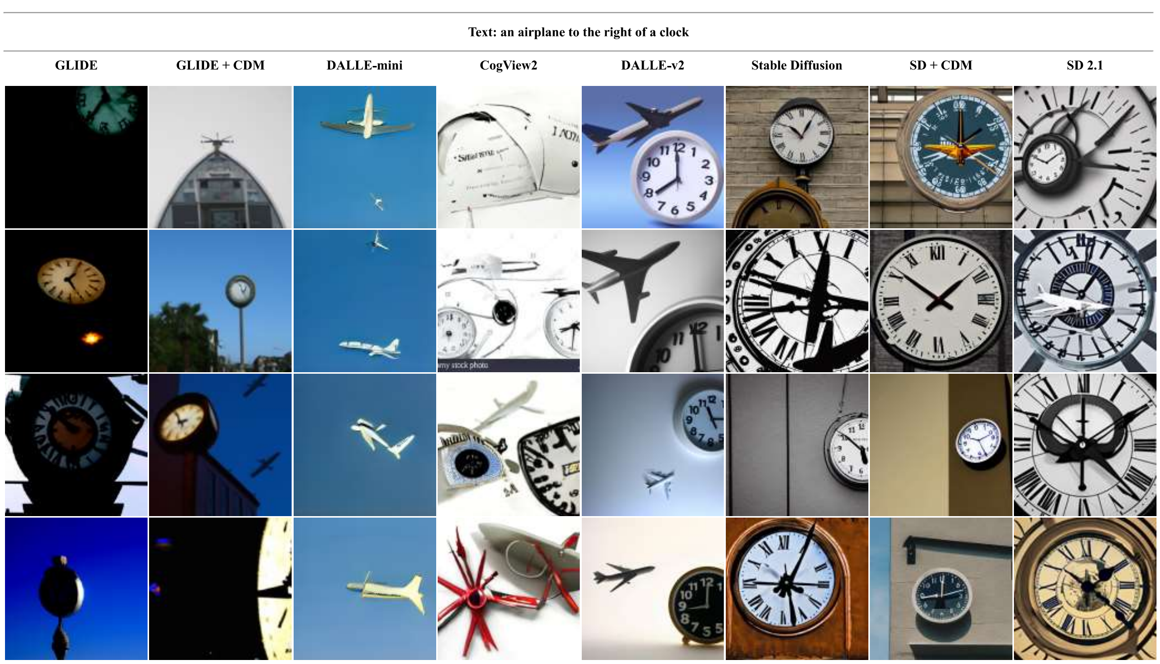



Figure 19: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.

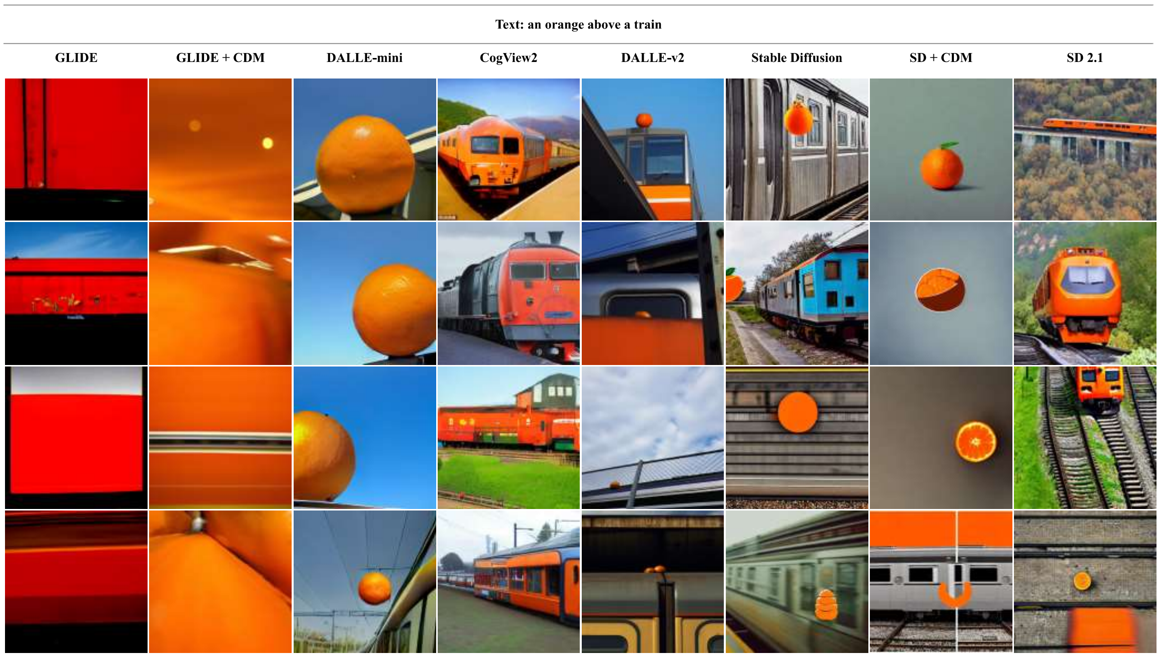



Figure 20: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.

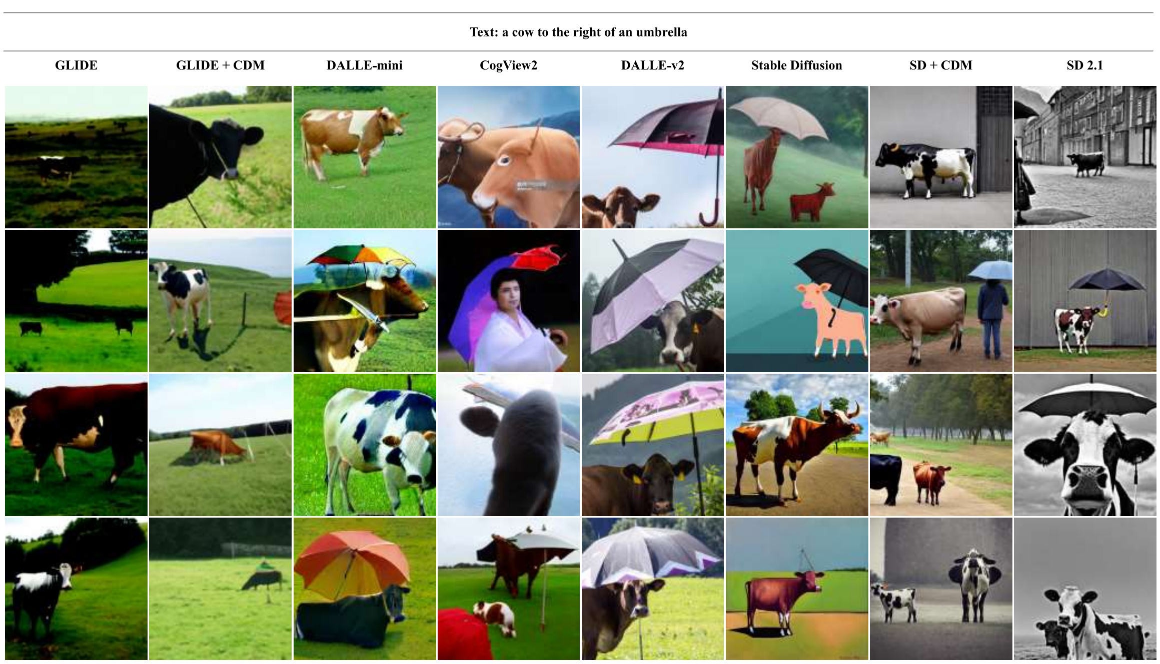



Figure 21: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.

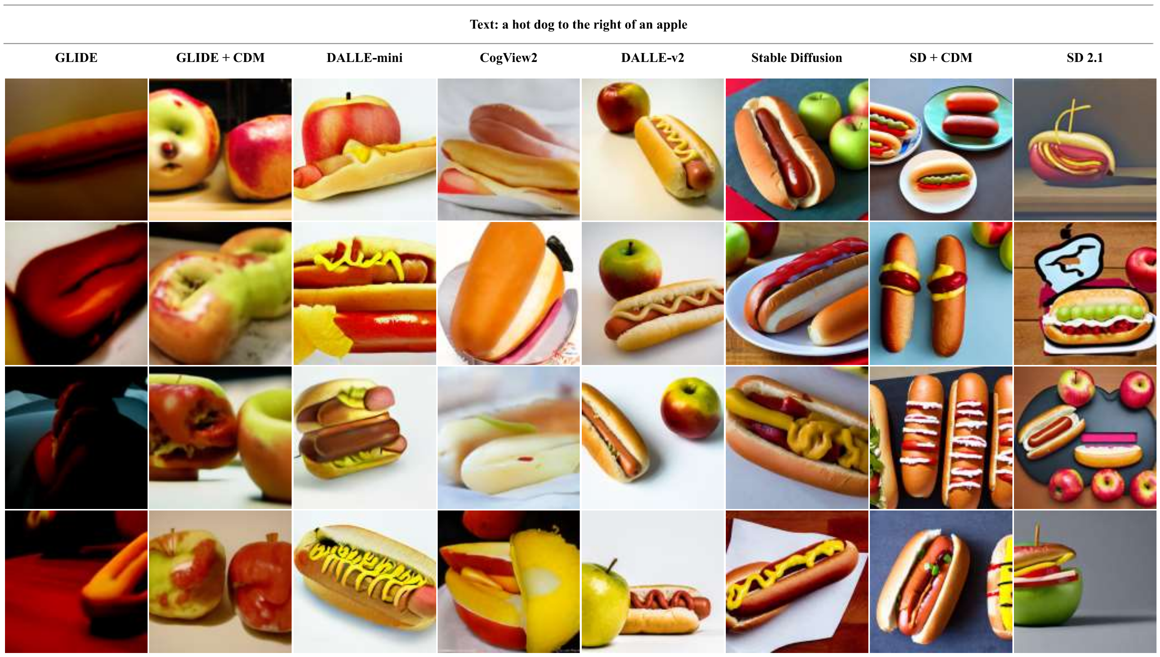



Figure 22: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.

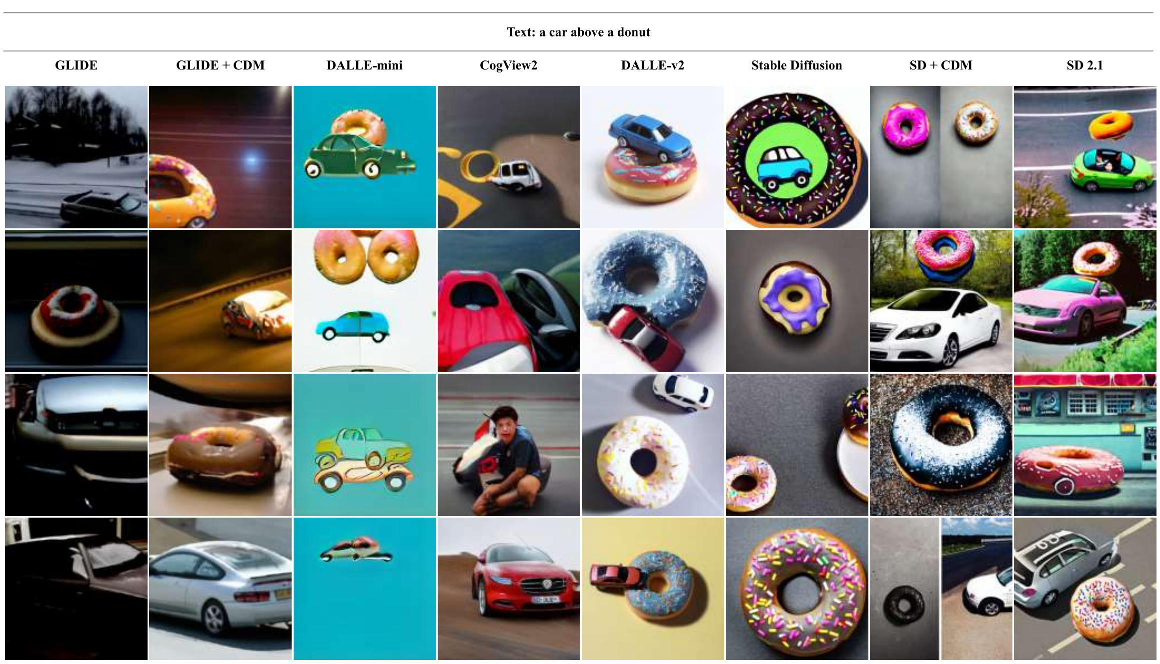



Figure 23: Illustrative examples of images generated by each of the 8 benchmark models using text prompts (top row) from the $SR_{2D}$ dataset.