# Diffusion Models as Visual Reasoners

**Jason Lin**[*],[1] **Maya Srikanth**[*],[1]

[1]Stanford University
{jj0, msrikant}@stanford.edu

## Abstract

Diffusion models have demonstrated powerful generation capabilities over recent years, achieving impressive performance on visual tasks such as text-guided image generation, inpainting, denoising and photorealistic image synthesis at high resolutions. However, even state-of-the-art diffusion models like DALL-E 2 (8) still struggle with basic visual reasoning: for instance, they often incorrectly represent compositional relationships, object counts, and negations in their text-guided generations (10). In this paper, we compare GLIDE vs Stable Diffusion and offer the following contributions: 1. probe visual reasoning failure modes during diffusion generation, 2. create a text-image dataset (*GQA-Captions*) from scene graphs for the purpose of improving text-to-image generation compositionality and 3. assess whether finetuning on spatially focused datasets can improve the compositional correctness of diffusion model generations both quantitatively and qualitatively (human evaluation). We also discuss limitations in existing quantitative metrics for assessing spatial reasoning in diffusion model generations. Our evaluations suggest that finetuning on spatially robust text-image data positively correlates with compositional correctness in diffusion generations. See our dataset and in-progress code here.

## 1 Introduction

Diffusion models have generated groundbreaking photorealistic images in recent years, but constrained by their text-guided encoders pretrained on noisy text-image data, they lack compositional awareness when guided with complex multi-object, single-sentence prompts (30). Nascent work have explored reasoning capabilities in diffusion models by deconstructing prompts into constituent concepts to be modeled by composable energy-based models. To bring generative models out of the art domain, vast applications would benefit, e.g. robotic simulation, medical imaging, but require compositionally accurate generative models at scale to be safely deployed.

Several visual reasoning datasets have emerged over recent years, including VQA (visual question answering) (12), CLEVR-X (11), and GQA (13), Visual Genome (24): they generally contain some combination of images, visually grounded questions about objects in the images, correct answers with optional bounding boxes, and natural language explanations. These datasets have primarily been used to train vision-language models to spatially reason about natural images by outputting answers in response to a natural language question in the style of VQA. Existing approaches to train such models include preprocessing object proposals with a visual detector while fusing entities and questions in a graph (25), as well as learning a joint vision-language embedding through (self-supervised) pretraining on proxy compositional and spatial tasks (e.g. VinVL (18), Unicoder-VL and ReCLIP (19)).

While spatially-focused text-to-image generation datasets have recently emerged (1), (4), they were created with the intent of helping models learn visual representations for downstream evaluation on various question-answering tasks. As such, they often lack spatially robust captions which describe important compositional relationships within an image. Most visio-linguistics datasets to date are evaluated on language tasks such as coreference resolution or referring expression comprehension (19) where dense bounding box annotations are necessary. To the best of our knowledge, these visual reasoning datasets have not yet been used to improve the spatial fidelity of text-guided image generations, and there is no work focused on compositional awareness in diffusion models during training.

## 2 Related Work

Diffusion models are a promising class of deep generative models that outperform likelihood-based models (e.g. VAEs) and implicit generative models (e.g. GANs) across a variety of benchmarks (14), offering benefits like stable training and controllable generation in addition to superior quality. During training, Gaussian diffusion models implement a forward diffusion process with a Markov chain of latent variables by progressively perturbing a sample $x_0 \sim q(x_0)$ from the input data distribution with Gaussian noise until the input is unrecognizable (16). In the reverse diffusion process, diffusion models learn the denoising distributions with a neural network, typically a U-Net or Transformer (16). Diffusion models can be trained to

[*]equal contribution.

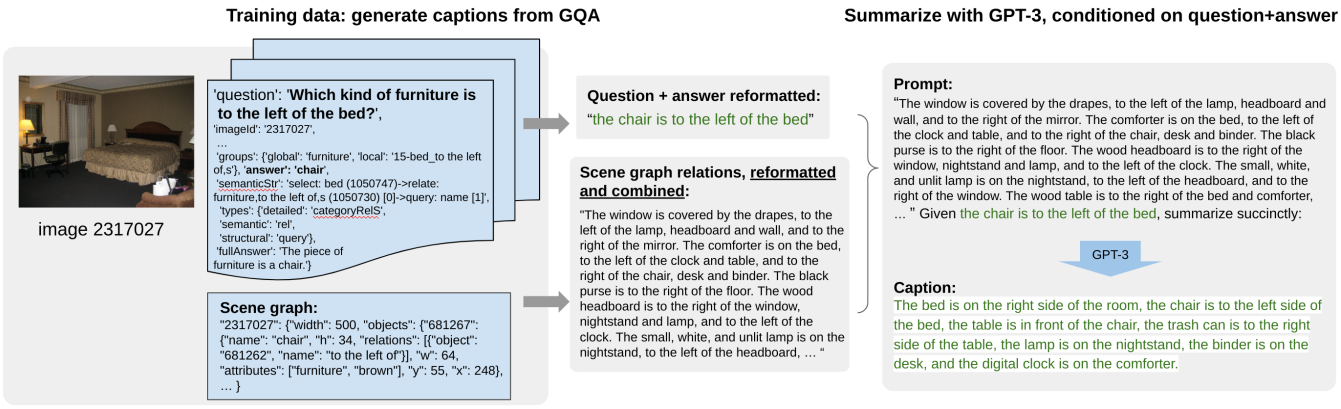


Figure 1: Our method for creating a compositionally-aware text-image dataset consists of two parts: parsing GQA scene graphs, then prompting a language model (GPT-3 (2)) along with a reformatted question + answer string in the form of "<long caption>" Given <question+answer>, summarize succinctly:

generate images conditional on various modalities (typically natural language), using gradients from a pretrained seed model to explicitly guide generation (classifier guidance) or by implicitly increasing the probability that the model conditions on relevant information during generation (classifier free guidance). (17)

**GLIDE vs. Stable Diffusion**: A major difference between GLIDE (20) and Stable Diffusion (9) (SD) is that SD diffuses in latent space before using a Variational Autoencoder to upsample into image space whereas GLIDE introduced a noisy CLIP for guided diffusion in pixel space. We compare most experiments between the two to assess the extent of improvement from finetuning.

**Compositional generation**: Concurrent to our work, (33) has shown that by viewing diffusion models as energy-based-models, conditional generations on conjunctions (AND) or negations (NOT) composed within a text prompt can synthesize semantically complex visual scenes. (33) introduced a CLIP-guided Contrastive Loss and Semantic matching loss to output a sentence direction in which to guide StyleGAN's latent generation. In (31), the authors proposed focal attention for object-to-object generation from a layout template with VQ-GANs. Another related work, (32) parses prompts into constituencies to be invidually embedded and conditioned in cross-attention during the sampling step without training.

**Dataset Construction**: To the best of our knowledge, there is no publicly available dataset of (caption, image) pairs where the captions summarize all objects and compositional relationships between the objects in the image (e.g. "there is a cat to the left of a dog"). To that end, we built a spatially enhanced dataset for text-to-image generation by extending GQA (13) in the style of Winoground's (caption, image) format. We also augmented entries with entity-specific questions, which are useful for downstream evaluation with an off-the-shelf VQA validator reference model (18). Our new dataset posits a challenging optimization task for any text-to-image diffusion model.

In this work, we aim to to understand the visual reasoning failure modes in diffusion models and determine whether finetuning on a spatially-focused text-image dataset can improve visual reasoning skills. Our contributions:

- Zero-shot experiments on visual reasoning capabilities of OpenAI's GLIDE-mini (20) and Stable Diffusion (9)
- GQA-Captions: we extend a subset of an existing dataset GQA (13) with images and corresponding scene graphs, into a text-image dataset with spatially-focused captions
- We finetune GLIDE-mini and Stable Diffusion on 2 spatially focused datasets, Winoground (1) and GQA-Captions and assess (both quantitatively and qualitatively) whether spatial reasoning improves.

We note that this paper describes ongoing work and results, detailing our creation of *GQA-Captions*, a new dataset to be released to motivate further study of compositions in diffusion models.

## 3 Method

### Dataset and task construction

In our study, we finetuned diffusion models on a novel extension to GQA (13), as well as a train subset of Winoground and evaluate on DrawBench (22) and Winoground (1). We provide details on how we constructed a *GQA-Captions* below.

### How do we generate captions from questions and scene graphs?

Prompting GPT-3 to summarize VQA-grounded questions and answers (in datasets like CLEVR-X and GQA), proves insufficient for generating a spatially descriptive caption about the images: the question-answer pairs provide sparse

coverage of objects in the image, and GPT-3 tends to imitate the question-answer format in the prompt. To create information-dense yet succinct captions which respect CLIP's 77-token text encoder limit, we parsed GQA scene graphs to generate natural language descriptions for each image. While information rich, the descriptions can be 400+ words long. Accordingly, we issue these descriptions, along with a final "task" sentence inspired by a spatially-focused GQA question associated with the image, to GPT-3 davinci for summarization (see Figure 3 for more details). We present an example of format of this prompt creation:

1. **Prompt Format**: Prompt GPT-3 davinci *(optimized for summarization) with a template of the form
   `<natural language description of GQA image>. Given <natural language description of objects in question corresponding to GQA image>, summarize succinctly:`
   This output retained most relevant spatial relations, although also removes adjectives (e.g. white, plastic straw → straw).

### Training Diffusion Models

We hypothesize that finetuning GLIDE (20) and Stable Diffusion's (9) decoders (while freezing their CLIP text encoders) on spatially focused text-image data can improve spatial-relationships in generated images. We experiment with different spatial reasoning datasets (GQA-Captions and Winground) as well as different training settings for GQA-Captions, including (1) finetuning for different amounts of time and (2) well as finetuning on less or more data. Our experiments are as follows:

1. GLIDE-mini no FT: OpenAI's GLIDE mini without additional finetuning
2. GLIDE-mini Wino FT: Finetuning GLIDE on Winoground for 20 epochs
3. GLIDE-mini GQA FT: Although we don't report results due to poor generation quality, we finetune GLIDE on 1200 GQA-Captions
4. SD no FT: Stable Diffusion checkpoint v1-5 without additional finetuning
5. SD Wino FT 100: Stable Diffusion additionally finetuned on a set of 600 Winoground examples for 100 epochs
6. SD GQA-1200 FT 8: Stable Diffusion finetuned on 1200 GQA-Captions examples for 8 epochs
7. SD GQA-4496 FT 11: Stable Diffusion finetuned no 4496 GQA-Captions examples for 11 epochs

Using 8-bit quantized Adam optimizer, batch size=1, FP16 mixed precision and gradient checkpoints, stable diffusion is able to fit in a RTX 3090 GPU and takes 6 hours to train 100 epochs.

### Metrics

Commonly used automated metrics for evaluating text-to-image generation include FID (Frechet Inception Distance) score, which focuses on higher-level image semantics, as well as retrieval-based (CLIP-)R-Precision (CRP), which assesses whether outputs are well-conditioned on the natural language prompt (23). Since FID isn't designed to penalize differences in spatial arrangement, we report (CRP) figures for our experiments. We find that CRP correlates moreso with human judgement than FID in terms of the compositional accuracy in generated images. In Table 1, we provide a quantitative comparison of FID and CRP which aligns with the notion that CRP is the better alternative. However, we acknowledge that CRP has its own weaknesses: (1) it is high variance because it relies on sampling captions from an arbitrary evaluation corpus, (2) given some generated image, it produces a ranking of captions in the evaluation corpus, and may thereby inflate scores if all evaluation captions are sufficiently distinct from one another, and (2) it is by design limited by the CLIP encoder's capabilities to represent spatial words in latent space, the same limitation which hinders CLIP-guided diffusion models like GLIDE-mini and Stable Diffusion.

| | **Groundtruth** | **SD v1-5** | **Wino FT** |
|---|---|---|---|
| FID unnorm. | - | 367 | 315 |
| CRP (pretrained) | 61.27 | **67.97** | 66.71 |
| CRP (Wino-FT) | 54.6 | 53.67 | **56.33** |

Table 1: Comparing FID score vs. CLIP-R-Precision (CPR) using groundtruth image as references, averaged over 800 Winoground caption-images. CPR correlated better than FID in overall semantic and compositional alignment, especially after finetuning the metric on Winoground test set.

Note that SD generations on the held-out portion of Winoground (after finetuning SD on Winoground-train) receive a marginally higher CRP score than ground-truth Winoground images, which indicates that CRP is an imperfect metric (see Table 1, row *CPR (pretrained)*). The row *FID unnorm* in Table 1 indicates that FID on Winoground is higher for baseline SD compared to SD finetuned on Winoground, which doesn't align with human judgement on the compositional correctness of SD generations after finetuning on Winoground. We also show in Table 1 that finetuning CRP's underlying CLIP encoder on Winoground-train results in figures which correlate moreso with human judgement on compositional awareness before and after finetuning SD on Winoground(see row *CRP (Wino-FT)*).

## 4 Results

### Zero-shot baseline

To assess whether diffusion models can capture compositional relationships in a zero-shot setting, we prompted GLIDE mini (GM) (20) and Stable diffusion v1-5 (SD) (9) with various Drawbench captions. Figure 2 shows GM is qualitatively worse than SD both in terms of generation quality and compositional accuracy. While we ran most of our experiments on both models, we show finetuning metrics on SD as qualitatively, GM lacked global coherence despite divesity across prompts and more consistent generations across repetitions. Further, we observed a trade-

*OpenAI Playground: https://beta.openai.com/playground

off between realism and compositional accuracy for several prompts: increasing the number of entities (i.e. three cats vs. one cat) in GLIDE's prompt compromises generation quality and realism. Even though SD is qualitatively better than GM, prompting the model to generate more than 3 objects (e.g. "4 cats") leads to similar drops in quality, i.e. incomplete portraints. We prompted stable diffusion to generate 800 images from Winoground's (1) captions and 70 images from Drawbench prompts. Generated images are available [†]here.

Figure 2: Generation for the prompt "A cat to the left of a dog", for GLIDE mini (**top**) and Stable Diffusion (**bottom**). Both models struggle to capture both entities (dog and cat) and their spatial relationship.

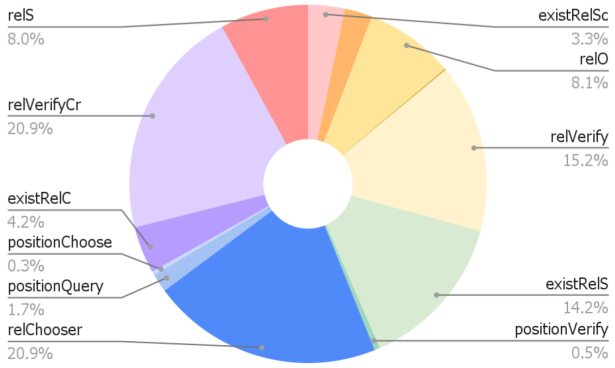


Figure 3: **GQA Question Types Used in GPT-3 Prompts.** To focus on a subset of scene objects during summarization, we append a "task" sentence to the natural language scene description parsed from GQA before issuing the prompt to GPT-3. This task sentence instructs to "succinctly summarize" the scene description conditioning on objects relevant to the structurally-aware groundtruth GQA questions. Distribution of question types are shown in the pie chart. "rel_" types often probe the spatial relationships between 2+ objects, those starting with "position" ask about the position of an an object relative to other objects, and "exist" types probe whether an object exists or doesn't exist within the scene.

[†] https://bit.ly/3tKnaPm

## Human Evaluation

Fig. 4 shows image generations using SD finetuned on either Winoground or GQA-Captions for different number of epochs compared to a vanilla pretrained SD checkpoint. Visually, all generations indicate that finetuning on 600 Winoground for 100 epochs or on 1200 GQA-Captions examples for 100 epochs can improve compositionality in text-guided generations. Although finetuning on spatially focused datasets can improve compositionality, we observe that this improved spatial robustness is accompanied by a drop in generation quality. We also note that training for less epochs on higher quality, smaller data like Winoground can lead to performance that rivals or outstrips performance after training on a larger, noisier dataset.

To further characterize the degree of compositional improvement, we need a robust signal alternative to high-variance, approximate metrics like CLIP-R precision. Accordingly, we provide a human evaluation on SD outputs for 39 DrawBench prompts (22) held-out from our training process. We selected prompts which focus on the tasks of *object counting* and *positional relationships*.

We compile results from two graders after asking them to (1) determine whether a generated image is compositionally "close" to the associated prompt and (2) rank their preferences for the generated images based on quality factors like realism, proportionality of the objects in the image, and subjective aesthetic factors. See Figure 5 and Figure 6. Our graders had high agreement such that our aggregated scores do not represent an average of two extremes.

**Findings.** Our human evaluation results indicate no finetuning of the pretrained Stable Diffusion checkpoint (SD no FT) results in the highest quality images in terms of aesthetics and realism, followed closely by SD finetuned on Winoground for 200 epochs (SD Wino FT 200). SD finetuned on a GQA-Captions subset of 1200 examples for 8 epochs (SD GQA 1200 FT 8) achieves comparable quality with SD finetuned on a GQA-Captions subset of 4496 examples for 11 epochs (SD GQA 4496 FT 11). Finally, SD finetuned on the 1200 GQA-Captions subset for 100 epochs (SD GQA 1200 FT 100) performed worst in terms of quality. Both our quantitative and qualitative results indicate that finetuning diffusion models on spatially-robust text-image datasets can lead to improved compositionality in image generations, potentially at the cost of quality and realism.

## 5 Future Work

**Dataset.** We continue to scale up our caption generation to more images (we have 4496 captions currently). We hope to also employ heuristics to filter out low-quality images (perhaps by testing how well a VQA model can answer questions about the image), and captions that don't mention essential objects in the images. Accompanied with finetuning analyses on the pruned dataset and potentially augmenting with CLEVR-X scene-caption pairs (11), we hope to open-source it for the ML community.

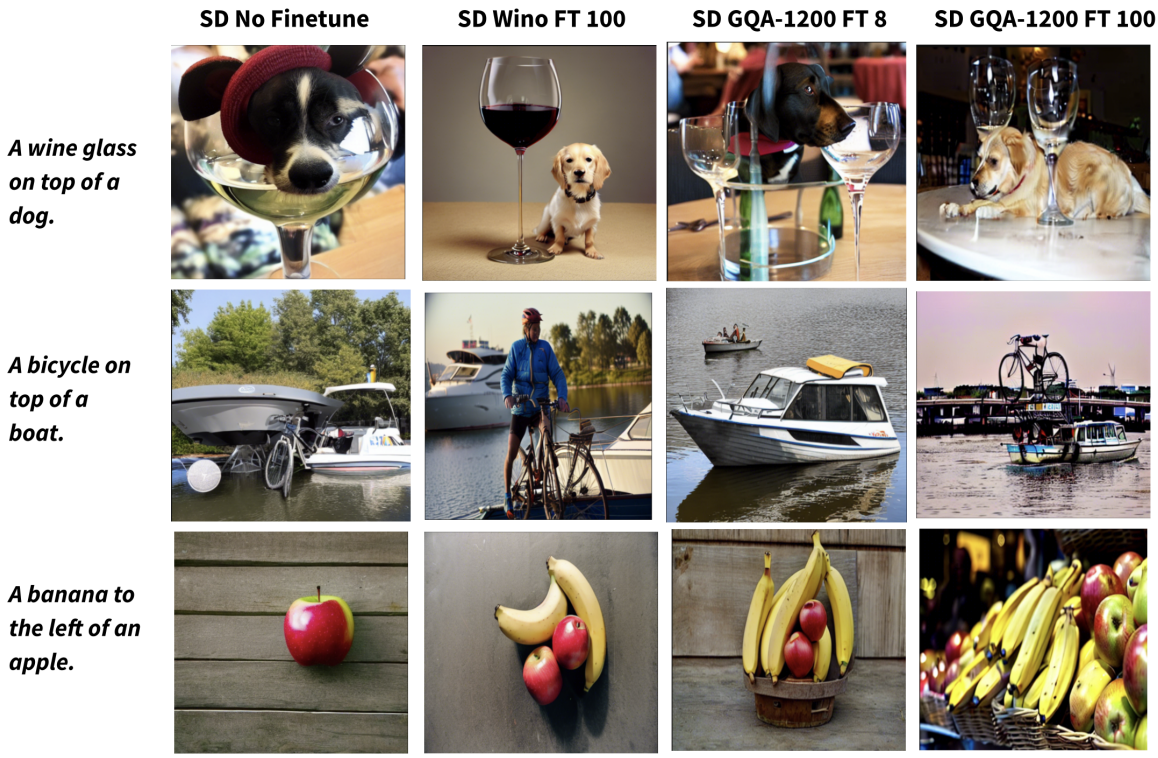


Figure 4: **Qualitative Results on SD Winoground FT and SD GQA FT**. Here, we show image generations for some drawbench prompts across SD No Finetune (stable diffusion v1-5 without finetuning, SD Wino FT 100 (SD finetuned on 600 winoground examples for 100 epochs), SD GQA-1200 FT 8 (SD finetuned on a subset of 1200 GQA-Captions example for 8 epochs),and SD GQA 1200 FT 100 (SD finetuned on 1200 GQA-Captions for 100 epochs).

| | SD v1-5 | SD Wino FT | SD GQA-1200 FT-9 | SD GQA-1200, FT-100 | SD GQA-4496 FT-11 |
|---|---|---|---|---|---|
| Winoground | 53.7 (4.32) | **56.3** (6.16) | 52.66 (6.38) | 51.39 (6.43) | 52.78 (7.68) |
| Drawbench | 50. (25.46) | 50. (12.42) | 55.56 (13.61) | **61.11** (12.42) | 52.78 (12.11) |

Table 2: CLIP-R-Precision for Stable Diffusion evaluated on Winoground and Drawbench benchmarks. Numbers in parenthesis are standard deviation. Results show finetuning on spatially focused data improves our metric on Winoground.

**Conditional Diffusion.** In addition to finetuning the U-Net in Stable Diffusion, we can leverage structured parsing for implicit guidance in the cross-attention layers during text-guided conditioning (32). Assuming the image content and layout can be disentangled by a separation of attention maps and values, we can apply a constituency parser to extract all Noun Phrases (NP) $N_i, i = 1, ..., k$ in a given prompt, encoded by the CLIP-text encoder individually.

**Finetuning.** Explore zero-initialization (26) in finetuning to approximate low-rank optimization and avoid overfitting (or catastrophic forgetting), and multiple instance learning (MIL)(27) for datasets with many-to-many image-questions. Another direction of work will explore automated prompt construction for other datasets in the absence of entity relations (NLVR2 (21)) and further set up a meta-training task for few-shot evaluation.

**Robust Evaluation.** Repeated diffusion sampling generates different images where some may match the composition of the prompt better than others. As such, we could select the best CLIP-R-Precision score out of $K$ SD generations for a given prompt. Another direction is to use Object Detection to Localize Objects with a granular spatial loss that is class-agnostic and amenable to lower-quality image generations or images with ambiguous or hybrid objects.

**Acknowledgement** We would like to thank our project mentor Kyle Hsu who helped refine the direction of our investigation. We also thank Chelsea Finn for sharing insights on few-shot adaptation and compositionally robust learning literature.

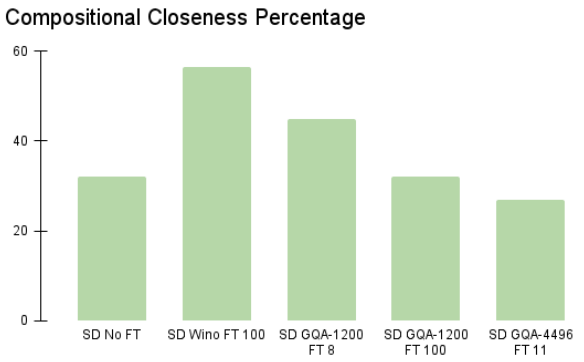


Figure 5: **Compositional Closeness Percentage.** We report the average percentage of prompts for which the generated image is marked as compositionally close to the prompt, averaged across the grader compositional closeness scores (binary, 1 if an image is close, 0 if it isn't)

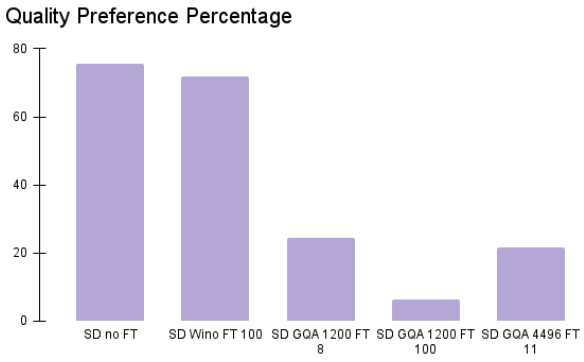


Figure 6: **Quality Preference Percentage. Quality Preference Percentage.** We report the percentage of times both graders ranked a generated image from one of our stable diffusion variants as their first or second preference out of all 5 options, across all 39 prompts.

## References

[1] Thrush, Tristan et al. "Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality." 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022): 5228-5238.

[2] Brown, Tom, et al. "Language models are few-shot reasoners." Advances in neural information processing systems 33 (2020).

[3] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models[J]. arXiv preprint arXiv:2205.01068, 2022.

[4] Liu, Fangyu, Guy Emerson, and Nigel Collier. "Visual Spatial Reasoning." arXiv preprint arXiv:2205.00363 (2022).

[5] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[6] Park, Dong Huk et al. "Benchmark for Compositional Text-to-Image Synthesis." NeurIPS Datasets and Benchmarks (2021).

[7] Ho, J., Jain, A., Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. arXiv. https://doi.org/10.48550/arXiv.2006.11239

[8] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv. https://doi.org/10.48550/arXiv.2204.06125

[9] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. arXiv. https://doi.org/10.48550/arXiv.2112.10752

[10] Marcus, Gary, Ernest Davis, and Scott Aaronson. "A very preliminary analysis of DALL-E 2." arXiv preprint arXiv:2204.13807 (2022).

[11] Salewski, Leonard, et al. "CLEVR-X: A Visual Reasoning Dataset for Natural Language Explanations." International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. Springer, Cham, 2022.

[12] Antol, Stanislaw, et al. "Vqa: Visual question answering." Proceedings of the IEEE international conference on computer vision. 2015.

[13] Hudson D A, Manning C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. "Proceedings of the IEEE/CVF conference on computer vision and pattern recognition." 2019.

[14] Song, Yang. "Generative Modeling by Estimating Gradients of the Data Distribution". https://yang-song.net/blog/2021/score/ (2021).

[15] Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Springer, Cham, 2020.

[16] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in Neural Information Processing Systems 34 (2021): 8780-8794.

[17] Luo, Calvin. "Understanding diffusion models: A unified perspective." arXiv preprint arXiv:2208.11970 (2022).

[18] P Zhang, X Li, X Hu, J Yang, L Zhang, L Wang, Y Choi, J Gao. "VinVL: Revisiting Visual Representations in Vision-Language Models." Proceedings of the IEEE/CVF Conference on Computer Vision." arXiv preprint arXiv:2101.00529 (2021).

[19] Subramanian, Sanjay, et al. "ReCLIP: A Strong Zero-Shot Baseline for Referring Expression Comprehension." arXiv preprint arXiv:2204.05991 (2022).

[20] Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." arXiv preprint arXiv:2112.10741 (2021).

[21] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

[22] Saharia, Chitwan, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." arXiv preprint arXiv:2205.11487 (2022).

[23] Park, Dong Huk et al. "Benchmark for Compositional Text-to-Image Synthesis." NeurIPS Datasets and Benchmarks (2021).

[24] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123.1 (2017): 32-73.

[25] Guo, Dalu, Chang Xu, and Dacheng Tao. "Bilinear graph networks for visual question answering." IEEE Transactions on Neural Networks and Learning Systems (2021).

[26] Zhao, Jiawei, Florian Schäfer, and Anima Anandkumar. "Zero initialization: Initializing residual networks with only zeros and ones." arXiv preprint arXiv:2110.12661 (2021).

[27] Tan, Reuben, et al. "NewsStories: Illustrating articles with visual summaries." European Conference on Computer Vision. Springer, Cham, 2022.

[28] Wah, Catherine, et al. "The caltech-ucsd birds-200-2011 dataset." (2011).

[29] Nilsback, Maria-Elena, and Andrew Zisserman. "Automated flower classification over a large number of classes." 2008 Sixth Indian Conference on Computer Vision, Graphics  Image Processing. IEEE, 2008.

[30] Liu, Nan, et al. "Compositional Visual Generation with Composable Diffusion Models." arXiv preprint arXiv:2206.01714 (2022).

[31] Yang, Zuopeng, et al. "Modeling Image Composition for Complex Scene Generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[32] Feng, Weixi, et al. "Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis." arXiv preprint arXiv:2212.05032 (2022).

[33] Liu, Nan, et al. "Compositional Visual Generation with Composable Diffusion Models." arXiv preprint arXiv:2206.01714 (2022).