REVISION: Rendering Tools Enable Spatial Fidelity in Vision-Language Models

Agneet Chatterjee[1(✉)], Yiran Luo[1], Tejas Gokhale[2], Yezhou Yang[1], and Chitta Baral[1]

[1] Arizona State University, Tempe, USA
agneet@asu.edu
[2] University of Maryland, Baltimore County, College Park, USA

Abstract. Text-to-Image (T2I) and multimodal large language models (MLLMs) have been adopted in solutions for several computer vision and multimodal learning tasks. However, it has been found that such vision-language models lack the ability to correctly reason over spatial relationships. To tackle this shortcoming, we develop the REVISION framework which improves spatial fidelity in vision-language models. REVISION is a 3D rendering based pipeline that generates spatially accurate synthetic images, given a textual prompt. REVISION is an extendable framework, which currently supports 100+ 3D assets, 11 spatial relationships, all with diverse camera perspectives and backgrounds. Leveraging images from REVISION as additional guidance in a training-free manner consistently improves the spatial consistency of T2I models across all spatial relationships, achieving competitive performance on the VISOR and T2I-CompBench benchmarks. We also design RevQA, a question-answering benchmark to evaluate the spatial reasoning abilities of MLLMs, and find that state-of-the-art models are not robust to complex spatial reasoning under adversarial settings. Our results and findings indicate that utilizing rendering-based frameworks is an effective approach for developing spatially-aware generative models. Code and data available at : https://github.com/agneet42/revision.

Keywords: Text to Image · Spatial Relationships · Rendering Graphics

1 Introduction

Generative vision-language models [36,44] represent a significant step towards developing multimodal systems that bridge the gap between computer vision and natural language processing. Text-to-image (T2I) models [5,38] convert

A. Chatterjee and Y. Luo—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73404-5_20.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15088, pp. 339–357, 2025.
https://doi.org/10.1007/978-3-031-73404-5_20

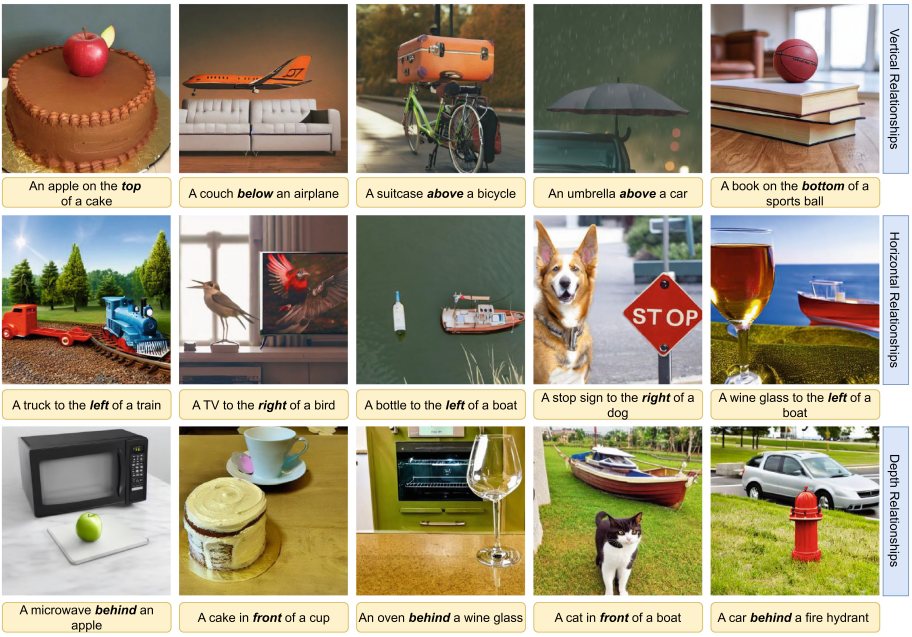


Fig. 1. Text-to-Image models struggle to generate images that faithfully represent the spatial relationships mentioned in the input prompt. We develop REVISION, an efficient rendering pipeline that enables a training-free and guidance-based mechanism to address this shortcoming. Our method results in improvements in spatial reasoning for T2I models for three dimensional relationships demonstrated by consistently higher scores on VISOR and T2I-CompBench benchmarks.

text prompts to high-quality images, while multimodal large language models (MLLMs) [25,48] process images as inputs, and generate rich and coherent natural language outputs in response. As a result, these models have found diverse applications in robotics [45], image editing [17], image-to-image translation [31], and more. However, recent studies [20] and benchmarks such as DALL-Eval [8], VISOR [15], and T2I-CompBench [18] have found that generative vision-language models suffer from a common mode of failure – their inability to correctly reason over spatial relationships.

We postulate that the lack of spatial understanding in generative vision-language models is a result of the lack of guidance from image-text datasets. Compared to T2I models, graphics rendering tools such as Blender allow deterministic and accurate object placement, but are limited by their lower visual detail and photorealism and do not have intuitive workflows such as T2I models where users can generate images by simply typing a sentence. To get the best of both worlds, in this work, we develop REVISION, a Blender-based image rendering pipeline which enables the synthesis of images with 101 3-dimensional object (assets), 11 spatial relationships, diverse backgrounds, camera perspectives, and lighting conditions. REVISION parses an input text prompt into assets

and relationships and synthesizes the scene using Blender to exactly match the input prompt in terms of both objects and their spatial arrangement.

In a *training-free* manner, we leverage images from REVISION as additional guidance for existing T2I methods to their ability to generate spatially accurate images, and demonstrate improved performance on VISOR and T2I-CompBench benchmarks. We evaluate (i) the impact of utilizing diverse backgrounds from REVISION, (ii) the trade-off between controllability and photo-realism and (iii) the added generalization to complex prompts achieved by leveraging REVISION. For a holistic study, we introduce an extension to the VISOR benchmark, to include evaluation of depth relationships (in front of/behind).

To assess the spatial and relational reasoning abilities of MLLMs, we also create the RevQA benchmark. We construct 16 diverse question types and their adversarial variations consisting of negations, conjunctions, and disjunctions. We perform holistic evaluations on 5 state-of-the-art MLLMs and discover significant shortcomings in their ability to accurately address complex spatial reasoning questions. These models also demonstrate a lack of robustness to adversarial perturbations, leading to a substantial decline in their performance.

The key contributions and findings are summarized below:

- We develop the REVISION framework, a 3D rendering pipeline that is guaranteed to generate spatially accurate synthetic images, given an input text prompt. An extendable framework, REVISION currently accommodates 100+ assets across 11 spatial relationships and 3 diverse backgrounds, and support for multiple lighting conditions, camera perspectives, and shadows.
- We present an approach that utilizes images from REVISION in an efficient training-free manner, which results in improved spatial reasoning across multiple benchmarks. Controlled experiments, ablations, and human studies reveal consistent improvements in generating images corresponding to the spatial relationships in the input prompt (as shown in Fig. 1).
- We introduce the RevQA question-answering benchmark to evaluate spatial reasoning abilities of multimodal large language models. Our experiments reveal the shortcomings of state-of-the-art MLLMs in reasoning over complex spatial questions and their vulnerability to adversarial perturbations.

2 Related Work

Generative Models for Image Synthesis. Image generation and synthesis methods have advanced rapidly, progressing from early approaches such as generative adversarial networks (GAN) [16], variational auto-encoders (VAE) [42], and auto-regressive models (ARM) [6], to contemporary text-to-image models including Stable Diffusion [38] and DALL-E [35]. GLIDE [30] adopts classifier-free guidance in T2I and explores the efficacy of CLIP [34] as a text encoder. Compared to GLIDE, Imagen [39] adopts a frozen language model as the text encoder, reducing computational overhead, allowing for usage of large text-only corpus. Multiple variants of T2I models have been developed by leveraging

T5-based text encoder [5], T2I priors [32,37], reward-based fine-tuning [18] and developing refiner models [33] for improved image-text alignment.

Controllable Image Generation for Spatial Fidelity. To achieve better control over diffusion-based image synthesis, multiple methods have been proposed. ReCo [49], GLIGEN [21], Control-GPT [53], Composable Diffusion [26] and ConPreDiff [47] all develop training-based methods to provide additional conditioning for T2I models. SPRIGHT [3] introduces a spatially-focused large-scale dataset, by re-captioning 6 million images from existing vision datasets and demonstrate performance gains through an efficient training methodology. Test-time adaptations have also been proposed - (i) Layout Guidance [7] restricts specific objects to their bounding box location through the modification of cross-attention maps; however it relies on bounding box annotations, (ii) LayoutGPT [12] and LLM-grounded Diffusion [23] leverage large language models (LLMs) to generate layouts and bounding box co-ordinates and, (iii) RealCompo [54] combines multiple generative models for better spatial control. By developing an annotation-free cost-efficient framework we overcome the shortcomings of existing methods through REVISION.

Synthetic Images for Vision and Language. The flexibility and control provided during creation of synthetic images has led to various visuo-linguistic evaluation benchmarks using rendering tools. CLEVR [19] pioneered the utilization of synthetic objects in simulated scenes for visual compositionality reasoning. Many variants of CLEVR such as CLEVR-Hans [43], CLEVR-Hyp [41], Super-CLEVR [22], and CLEVRER [50] probe multiple facets of multimodal understanding with synthetic images and videos. PaintSkills introduced in DALL-EVAL [8] is an evaluation dataset that measures multiple aspects of a T2I model, which includes spatial reasoning, image-text alignment and social biases.

Evaluation of Multimodal LLMs. Multiple benchmarks have been proposed that evaluate reasoning capabilities of MLLMs. MMBench [27] evaluates models across 20 different dimensions, for a total of 2974 evaluation instances. The distinctive abilities of MLLMS to differentiate between coarse and fine-grained vision tasks is explored by MME [13] with images sourced from COCO. A limitation across all these benchmarks is that they collect instances from common VL datasets, increasing risk of data leakage *and* do not evaluate spatial relationships at scale. RevQA fills this gap by developing a diverse set of synthetic and scalable image-question pairs for a holistic evaluation.

3 The REVISION Framework

REVISION (Fig. 2) is a rendering-based framework for generating spatially accurate images from an input prompt. Given a prompt, we generate an image in Blender, where the two object 3D models and the camera view are situated according to the spatial relationship derived from the prompt. The components of REVISION are described below.

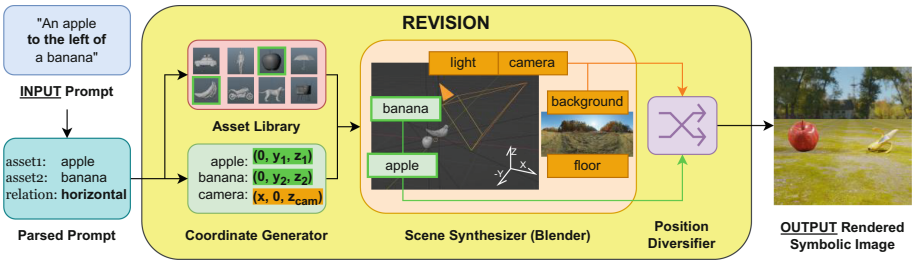


Fig. 2. REVISION parses a prompt into assets (objects) and the spatial relationship between them and synthesizes a symbolic image in Blender, placing the respective object assets at coordinates corresponding to the parsed spatial relationship.

Table 1. Spatial relationships in REVISION and their rules for the Coordinate Generator. The objects are positioned from the camera's perspective.

Relation	Spatial Phrases	Coordinate Constraints	Distance (m)
Horizontal	**to the left, to the right**	X and Z are 0.	[1, 1.5]
Vertical	**above, below, top, bottom**	X and Y are 0.	[0.75, 1]
Near	**near, next to, on the side of**	Z is 0.	[0.75, 1]
Depth	**in front of, behind**	Z is 0. X_{obj1}=-X_{obj2}.	[1, 1.5]

The **Asset Library** includes a large human-inspected collection of 3D models of realistic objects with variations in texture and shape. Given an object name, the Asset Library randomly selects a matching asset rescaled to fit into a $1m$ cube to ensure that they are sufficiently visible in the final output. The Asset Library features 101 distinct classes of objects, 80 of which are from MS-COCO [24]. Each object class is associated with 3 to 5 royalty-free 3D model assets from sketchfab.com, with a total of 410 3D models. REVISION includes 3 background panoramas (Indoor, Outdoor, and White) from polyhaven.com and a corresponding textured floor asset from Sketchfab.

The **Coordinate Generator** deterministically generates 3D coordinates for the objects and the camera, given the names of the objects and the spatial relation extracted from the prompt. As shown in Table 1, REVISION supports four categories of spatial relationships between objects. In our coordinate frame, the X-, Y-, and Z-axis represent depth, horizontal, and vertical relationships respectively. To ensure that the objects are visible and the spatial relationship is obvious from the camera's view, the coordinate values for the objects on all three axes are confined within the range of $[-1\,\mathrm{m}, 1\,\mathrm{m}]$. The camera is placed at $x = 5\,\mathrm{m}$ with its view always facing the origin point. The camera is at $z = 2.5$ for depth relationships and at $z = 1.5m$ otherwise.

The **Scene Synthesizer** assembles a 3D scene consisting of six main components: a camera, a light source, background, floor, and two objects. The two object assets and the camera are placed at their respective coordinates deter-

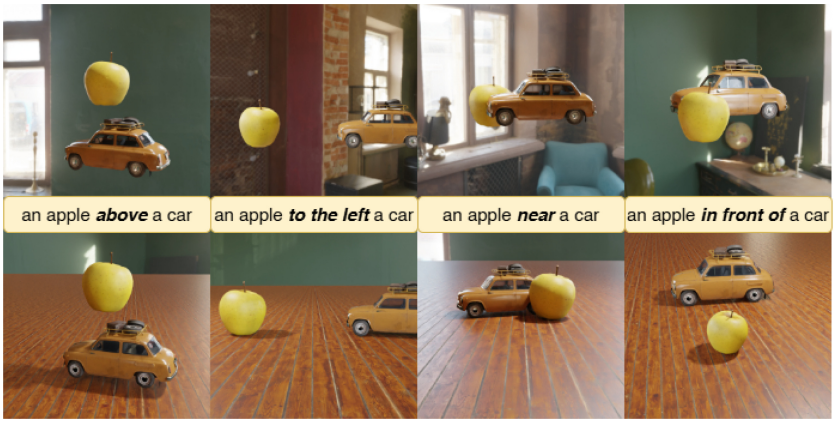


Fig. 3. Outputs from the REVISION rendering pipeline for 4 spatial relationships types for identical assets, with (**bottom**) and without a floor (**top**).

mined by the Coordinate Generator. Then the background asset, which is a 360-degree panorama image (modeled as a large sphere), is centered at the origin. The light source is added to a random position sufficiently higher than all objects in the scene. To prevent objects from appearing to float, the floor asset, a textured hyperplane orthogonal to the Z-axis, is positioned beneath the object asset with the lowest vertical coordinate. This floor placement also enables the object assets to cast shadows, enhancing the realism of the rendered image.

The **Position Diversifier** (Fig. 3) ensures diversity in object orientations, background, and the camera angles every time REVISION is invoked. The background is rotated along the Z-axis, giving us a large number of static background options. In order to further diversify the perspective sizes and tilts of the object assets within the camera's view, we add random jitter to the position and orientation of the camera. We also add random small rotations to the objects along the Z-axis and vary the distance between the objects so that they are not always symmetric around the origin. See Supplementary Materials for more details.

4 Improving Spatial Fidelity in T2I Generation

4.1 Training-Free Image Generation with REVISION

Given an input prompt (T), we first generate a spatially accurate reference image ($x^{(g)}$) leveraging our REVISION pipeline. We then perform training-free image synthesis to generate an image I, i.e. $\phi(I|x^{(g)}, T)$, where ϕ is a T2I model. We reformulate the standard text-to-image pipeline into an image-to-image pipeline, conditioned by text, as shown in Fig. 4.

Standard diffusion methods such as Stable Diffusion (SD) generate an image by iteratively de-noising a Gaussian noise vector. Stochastic Differential Editing (SDEdit) [28], on the other hand, starts from a guide image ($x^{(g)}$, in our case),

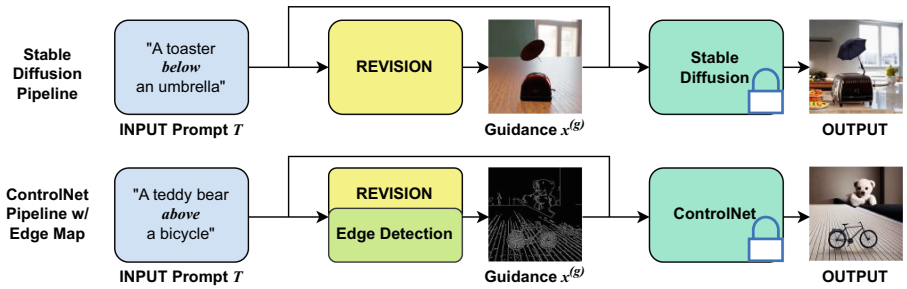


Fig. 4. Given a user-provided input prompt T, we generate a corresponding synthetic image $x^{(g)}$ using REVISION. With input prompt T and guidance $x^{(g)}$, we perform training-free image synthesis based on existing T2I pipelines such as Stable Diffusion or ControlNet to obtain a spatially accurate image.

Table 2. The incorporation of REVISION as a guiding framework significantly enhances the spatial reasoning performance of Stable Diffusion (SD) models. Results highlighted in green represent scores achieved with images from REVISION.

Method	OA (%)	VISOR (%)					
		uncond	cond	1	2	3	4
SD 1.4	29.86	18.81	62.98	46.60	20.11	6.89	1.63
+ REVISION	53.96	52.71	97.69	77.79	61.02	44.90	27.15
SD 1.5	28.43	17.51	61.59	44.27	18.12	6.28	1.35
+ REVISION	54.33	53.08	97.72	78.07	61.27	45.44	27.55
SD 2.1	47.83	30.25	63.24	64.42	35.74	16.13	4.70
+ REVISION	48.26	47.11	97.61	76.07	55.75	37.10	19.53

adds Gaussian noise to it, and denoises it to produce the synthesized image I. We use SDEdit within our Stable Diffusion pipeline and perform image generation guided by $x^{(g)}$. We also explore the ControlNet [51] backbone, which allows fine-grained control over SD. Using ControlNet allows us to address two key points: **a)** our reference images provide enough spatial information even when low-level features are extracted from them and, **b)** we can mitigate any attribute-related biases present in the assets.

4.2 Experimental Setup

We study the efficacy of REVISION on two widely accepted benchmarks for spatial relationship, VISOR [15] and T2I-CompBench [18], which have 25,280 and 300 spatial prompts, respectively. For each evaluation prompt in the respective benchmarks, we generate a corresponding image from our REVISION pipeline and perform training-free image generation as described in Sect. 4.1.

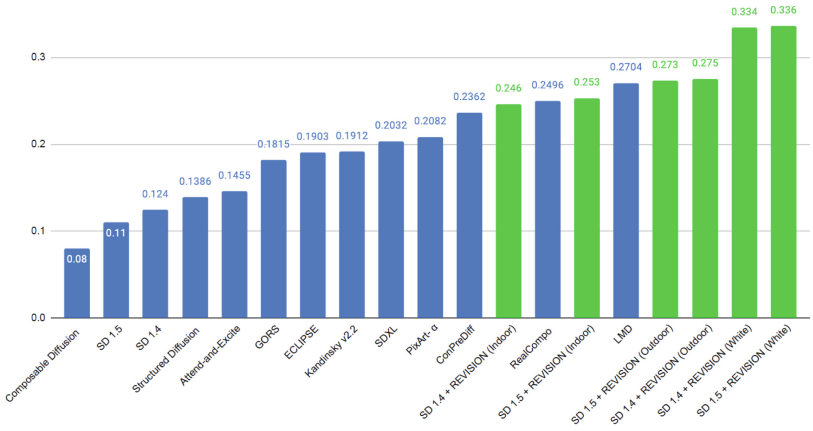


Fig. 5. Comparing the T2I-CompBench spatial scores of REVISION-based guidance (green) with other leading T2I models and methods (blue). (Color figure online)

Table 3. Results on the VISOR Benchmark. With REVISION, we consistently outperform existing T2I methods on the VISOR benchmark.

Method	OA (%)	VISOR (%)					
		uncond	cond	1	2	3	4
GLIDE [30]	3.36	1.98	59.06	6.72	1.02	0.17	0.03
DALLE-mini [10]	27.10	16.17	59.67	38.31	17.50	6.89	1.96
DALLE-v2 [35]	**63.93**	37.89	59.27	73.59	47.23	23.26	7.49
Layout Guidance [7]	40.01	38.80	95.95	-	-	-	-
Control-GPT [53]	48.33	44.17	65.97	69.80	51.20	35.67	20.48
Structured Diffusion [11]	28.65	17.87	62.36	44.70	18.73	6.57	1.46
Attend-and-Excite [4]	42.07	25.75	61.21	49.29	19.33	4.56	0.08
SD 1.4 + REVISION	53.96	52.71	<u>97.69</u>	77.79	61.02	44.90	27.15
SD 1.5 + REVISION	54.33	<u>53.08</u>	**97.72**	<u>78.07</u>	<u>61.27</u>	<u>45.44</u>	<u>27.55</u>
SD 2.1 + REVISION	48.26	47.11	97.61	76.07	55.75	37.10	19.53
ControlNet + REVISION	<u>56.88</u>	**55.48**	97.54	**78.82**	**62.93**	**48.58**	**31.59**

We leverage 3 variants of Stable Diffusion (SD), versions 1.4, 1.5, and 2.1 as our baseline models. For ControlNet, we use the canny edge-conditioned SD model. For holistic evaluations, we also report the Inception Score (IS) [40] where applicable. For all subsequent tables, the **bold** values denote the best performance while <u>underlined</u> values indicate the second-best performance.

4.3 Results and Analysis

Improvements over Baseline Models - We summarize our representative improvements over the baseline and existing methods, on the VISOR and

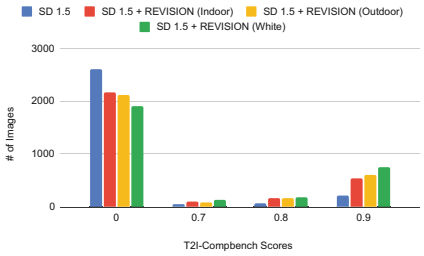


Fig. 6. REVISION improves T2I-CompBench Spatial Score (0 indicates missing objects, 1 denotes perfect object generation and spatial accuracy.)

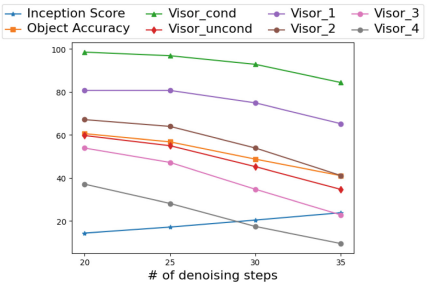


Fig. 7. Benchmarking the trade-off between spatial accuracy (VISOR) and Inception Score, achieved with REVISION.

Table 4. VISOR$_{cond}$ and Object Accuracy, split across relationship types. $\sigma_{\texttt{Vc}}$ and $\sigma_{\texttt{OA}}$ denote the respective metric's standard deviation w.r.t the relationships. Regardless of the spatial relation, REVISION enables T2I models to consistently produce spatially accurate images, a challenge faced by earlier approaches.

Method	VISOR$_{cond}$ (%)					Object Accuracy (%)				
	left	right	above	below	$\sigma_{\texttt{Vc}}$	left	right	above	below	$\sigma_{\texttt{OA}}$
GLIDE	57.78	61.71	60.32	56.24	2.46	3.10	3.46	3.49	3.39	0.18
DALLE-mini	57.89	60.16	63.75	56.14	3.29	22.29	21.74	33.62	30.74	5.99
DALLE-v2	56.47	56.51	60.99	63.24	3.38	**64.30**	**64.32**	**65.66**	**61.45**	1.77
Control-GPT	72.50	70.28	67.85	65.70	2.95	49.80	48.27	47.97	46.95	1.18
SD 1.4 + REVISION	97.53	97.45	**98.09**	97.66	0.29	52.42	52.11	56.93	54.38	2.22
SD 1.5 + REVISION	97.57	**97.53**	98.05	97.70	0.24	52.99	52.59	56.80	54.92	1.94
SD 2.1 + REVISION	**97.81**	97.46	97.91	97.28	0.30	46.70	47.94	49.70	48.71	1.27
ControlNet + REVISION	97.51	97.25	97.65	**97.72**	0.21	55.10	55.14	58.98	58.29	2.05

T2I-CompBench benchmarks in Table 2 and Fig. 5 respectively. The results in Table 2 are shown with reference images on a white background and # of denoising steps = 30. As shown in Table 2, we improve on all aspects of spatial relationships compared to our baseline methods. On SD 1.5, we achieve a **91.1%** improvement in Object Accuracy (OA) and a **58.6%** improvement on the conditional score. Specifically, we generate objects more accurately and achieve a high % of accuracy when spatially synthesizing them in the image. Interestingly, through REVISION, we increase the likelihood of consistently generating spatially correct images, as can be seen by the relatively high value of VISOR$_4$. On VISOR (Table 3), REVISION enables baseline Stable Diffusion models to consistently outperform existing methods, across all aspects. Compared to the best open-source model, Control-GPT, we achieve a Δ improvement of **17.69%**, **48.12%**, and **25.6%** on OA, VISOR$_{cond}$, and VISOR$_{uncond}$ respectively.

Fig. 8. Illustrative examples depicting the variation of generated images across the three variants of backgrounds in REVISION. For each pair, the image on the left is from REVISON and the image on the right is generated from the T2I model.

On T2I-CompBench (Fig. 5), we observe similar improvement trends across diverse backgrounds, with baseline models guided by REVISION achieving consistent performance gains on the benchmark. In addition to enhancing spatial accuracy, REVISION improves prompt fidelity by ensuring that images contain all objects mentioned in the input prompt (Fig. 6).

Consistent Performance Across Relationship Types - Across all spatial relationship types, REVISION achieve a consistently high performance score across the VISOR metrics as shown in Table 4; a shortcoming prevalent in other methods. For example, the largest deviation in VISOR_{cond} performance for ControlNet + REVISION is 0.21% between *left* and *below* relationships; in comparison Control-GPT deviates as much as 6.8% for the same.

4.4 Ablation Studies

Impact of Background - In Table 5, we enumerate the impact of the background types in the images from the REVISION pipeline and the downstream trade-off between VISOR performance and model diversity. Utilizing white back-

Table 5. The impact of the 3 background types in the REVISION pipeline on the VISOR benchmark. While best performance is achieved with a white background, diverse outputs are attained with the outdoor background type.

Model	Background	IS (↑)	OA (%)	VISOR (%)					
				uncond	cond	1	2	3	4
SD 1.4	White	16.16	53.96	52.71	97.69	77.79	61.02	44.9	27.15
	Indoor	19.11	48.53	45.12	92.97	74.82	53.79	34.78	17.09
	Outdoor	**20.16**	44.32	41.80	94.31	69.79	49.38	31.86	16.17
SD 1.5	White	16.27	**54.33**	**53.08**	**97.72**	**78.07**	**61.27**	**45.44**	**27.55**
	Indoor	19.11	48.77	45.28	92.85	74.93	53.96	34.77	17.47
	Outdoor	19.66	43.99	41.51	94.36	69.48	48.58	31.46	16.52
SD 2.1	White	12.79	48.26	47.11	97.61	76.07	55.75	37.10	19.53
	Indoor	11.52	31.08	29.37	94.50	59.80	33.96	17.40	6.34
	Outdoor	10.51	36.37	34.67	95.34	65.05	41.23	23.05	9.36

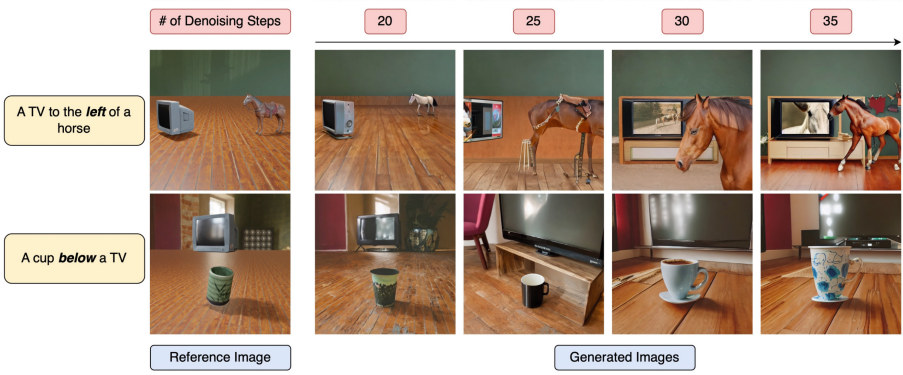


Fig. 9. Illustrative examples showing the trade-off between photo-realism and denoising steps, while maintaining generating spatially accurate images using REVISION.

grounds that exclusively feature the two objects in question minimizes potential distractions for the model. Conversely, when the model is presented with initial reference images incorporating indoor or outdoor backgrounds, it exhibits the capacity to identify and leverage *distractor* objects, resulting in the generation of diverse images. As shown in Fig. 8, all generated images maintain spatial accuracy, but noisier reference images result in greater diversity.

Controllability vs Photo-Realism - In this setup, we study the impact of the # of denoising steps and its trade-off with photo-realism. As shown in Fig. 7 that while the performance on VISOR deteriorates with additional # of denoising steps, it improves the model's ability to be more diverse and photo-realistic. In Fig. 9, we demonstrate that by utilizing REVISION, baseline models

Table 6. Comparing baseline methods against REVISION-guided image synthesis on depth relationships. DS denotes the # of Denoising Steps.

Setting	OA (%)	VISOR (%)				
		uncond	1	2	3	4
SD 1.5 (baseline)	41.52	27.15	60.12	31.82	13.19	3.47
+ REVISION (DS = 30)	**58.32**	29.62	64.74	34.53	14.99	4.22
+ REVISION (DS = 35)	52.05	**32.08**	**68.11**	**37.92**	**17.51**	**4.78**
+ REVISION (DS = 40)	47.43	30.46	65.50	35.99	15.74	4.64

can preserve their spatial coherency while iteratively demonstrating a higher degree of photo-realism, through more # of denoising steps.

4.5 Extending VISOR for Depth Relationships

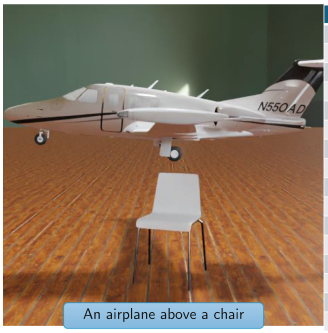
We further extend the VISOR benchmark for Depth relationships (in front of/behind). We utilize Depth Anything [46] for generation of depth maps and OWLv2 [29] for object detection. Given a T2I generated image I and its prompt T that contain two objects o_1, o_2, we obtain its depth map I_D using Depth Anything. We then retrieve the centroids detected for the two objects c_{o_1}, c_{o_2} using OWLv2. At these centroid coordinates, we acquire the depth values for the two objects from the depth map $I_D(c_{o_1}), I_D(c_{o_2})$. We check if the acquired depth values match the spatial relationship in the prompt, and evaluate similar to VISOR. As shown in Table 6, REVISION improves VISOR scores across all metrics and across multiple denoising steps.

4.6 Human Evaluations

To verify the generalizability of REVISION-based guidance on T2I models, we perform 2 distinct experiments and conduct human evaluations for validation. For each experiment, we independently sample 200 generated images and take the average scores across 4 workers. We also report unanimous (100%) and majority (75%) agreements between the workers for each experiment.

Prompts of Multiple Objects and Relationships - In this experiment, we generate reference images using prompts that include 2 spatial phrases and 3 objects, and use these images to guide T2I generation. Each generated image is evaluated for accuracy based on the input spatial prompts. We achieve an accuracy of 79.62% when at least 1 phrase is correctly represented in the image and 46.5% when both phrases are correctly represented. The unanimous and majority agreements among evaluators are 64.5% and 86.5%, respectively.

Out-of-Distribution Objects - We consider prompts containing exactly one object not found in the REVISION Asset Library. Given a prompt that mentions an OOD object, we find the semantically closest object (list in Supplementary



Question Type	Question	Answer
Simple Spatial	Is there an airplane ***above*** a chair?	Yes
Opposite Spatial	Is there a chair ***above*** an airplane?	No
AND	Is there a chair ***and*** an airplane?	Yes
OR	Is there a chair ***or*** an airplane?	Yes
NOT	Is there an airplane ***not above*** a chair?	No
Double Negative	Is there an airplane ***not below*** a chair?	Yes
Random AND	Is there an airplane ***and a hot dog?***	No
Random OR	Is there an airplane ***or a hot dog?***	Yes
Random Spatial	Is there an airplane ***above a hot dog?***	No
Random Combined AND	Is there an airplane above a chair **and** is there an airplane above a hot dog?	No
Random Combined OR	Is there an airplane above a hot dog **or** is there an airplane above a chair?	Yes
Adversarial AND	Is there a ***helicopter and*** a chair?	No
Adversarial OR	Is there a ***helicopter or*** a chair?	Yes
Adversarial Spatial	Is there a ***helicopter above*** a chair?	No
Adversarial Combined AND	Is there a ***helicopter above*** a chair ***and*** is there is an airplane ***above*** a chair?	No
Adversarial Combined OR	Is there a ***helicopter above*** a chair ***or*** is there is an airplane **above** a chair?	Yes

Fig. 10. The RevQA Benchmark. Using the REVISION pipeline, we generate spatially accurate images and formulate 16 question types from a given caption. We leverage these generated questions and image, benchmarking Multimodal Large Language Models in their abilities to reason over spatial relationships.

Material) in our library and use their corresponding image as guidance. For example, we generate an image of "a ***helicopter*** above a bicycle" by providing a reference image of "an ***airplane*** above a bicycle". An accuracy of 63.62% is found with an unanimous and majority agreement of 67% and 90.5%, respectively.

5 RevQA: A Spatial Reasoning Benchmark for MLLMs

We leverage the determinism of the REVISION pipeline to construct a new visual question answering benchmark (RevQA) for evaluating the spatial reasoning abilities of multimodal large language models.

Question Generation. The benchmark contains 16 types of yes-no questions for a REVISION-generated image, consisting of negations, conjunctions, and disjunctions, building on prior work on logic-based visual question answering [14]. Each question type evaluates a combination of spatial and logical reasoning abilities in multimodal large language models (MLLMs) (Fig. 10).

Among the 16 types, we incorporate *Random* and *Adversarial* types of questions to further evaluate the robustness and reliability of MLLMs using simple templated transformations. In *Random* types of questions, we replace an object (visible in the image) in the question with another randomly picked object from REVISION's Asset Library. For the *Adversarial* set of questions, we replace one of the objects with another that is semantically and visually close. In addition to benchmarking their robustness, these questions allow simultaneous evaluation of the fine-grained spatial perception and reasoning abilities of these models. To alleviate any order bias in instances which contain multiple questions (see *Combined* in Fig. 10), we randomly switch the order between them.

Evaluation Setup and Results. We sample *50k* image-question pairs and benchmark 5 open-source state-of-the-art MLLMs - LLaVA 1.5 [25], Fuyu-8B [2],

Table 7. Performances of 5 MLLMs across the 16 types of questions in RevQA. Most models perform worse than random (50%) when reasoning over Opposite Spatial relationships and Double Negative questions. All models have a significant ***drop*** in performance with Random/Adversarial questions, in comparison to their simpler versions.

Question Type	LLaVa 1.5	Fuyu-8B	InstructBLIP	LLaMA-Adapter 2.1	Qwen-VL Chat
Simple Spatial	**0.942**	0.702	0.834	0.579	0.940
Opposite Spatial	0.394	0.287	0.184	**0.419**	0.402
AND	0.935	0.887	**0.957**	0.858	0.889
OR	**0.995**	0.396	0.598	0.722	0.949
NOT	**0.926**	0.619	0.356	0.504	0.583
Double Negative	0.267	0.347	**0.665**	0.490	0.212
Random AND	0.934	0.308	0.675	0.616	**0.978**
Random OR	**0.925**	0.178	0.194	0.194	0.324
Random Spatial	**0.925**	0.370	0.686	0.790	0.919
Random Combined AND	0.116	0.502	0.627	**0.800**	0.567
Random Combined OR	**0.968**	0.536	0.414	0.003	0.506
Adversarial AND	0.661	0.184	0.542	0.641	**0.789**
Adversarial OR	**0.921**	0.188	0.443	0.156	0.685
Adversarial Spatial	0.559	0.335	0.777	**0.893**	0.615
Adversarial Combined AND	0.132	0.539	0.695	**0.805**	0.695
Adversarial Combined OR	**0.953**	0.456	0.386	0.003	0.254
Average	**0.720**	0.446	0.598	0.578	0.642

InstructBLIP [9], LLaMA-Adapter 2.1 [52] and Qwen-VL-Chat [1]. We instruct all models to generate binary responses and set the temperature = 0, to remove stochasticity in the generated responses.

We present our evaluation results in Table 7 and find that all models have a large gap in performance in reasoning over spatial relationships. While most models reason well over simple spatial relationships, they have a large performance drop when presented with the opposite spatial relationships. For example, LLaVA-1.5, the best performing model, has a 58.17% decrease in performance when probed with *simple* vs *opposite* spatial questions. This can be attributed to: ***(a)*** insufficient training data for rare object relationships, such as less instances of an "elephant *above* a person" than vice versa; ***b)*** the inability of vision encoders like CLIP to capture subtle semantic differences. MLLMs also struggle with negation, possibly because image captions do not capture enough negations; e.g. COCO Captions only contain***0.97%*** occurrences of *'not'*. All models significantly suffer when presented with questions that consist of *double negatives*, which evaluate the models' ability to reason of negations and spatial relationships in tandem. Furthermore, all models suffer under adversarial settings in comparison to their simpler counterparts; comparing LLaVA's performance for *AND* and *Adversarial Combined AND* questions, we find a 85.88% ($0.935 \rightarrow 0.132$) drop in performance. We also observe a larger decline in performance for *Adversarial* questions than for the *Random* set of questions hinting that while models independently perform well at object recognition and simple spatial relationships, combining them adversarially significantly reduces performance.

6 Conclusion

In this work, we introduce REVISION, a framework designed for training-free enhancement of spatial relationships in Text-to-Image models and RevQA, a benchmark to evaluate the spatial reasoning abilities of multimodal large language models. Our results demonstrate the effectiveness of leveraging 3D rendering pipelines as a cost-efficient approach for developing generative models with robust reasoning capabilities. REVISION is modular and can easily be extended to incorporate additional features, assets, and relationships. We hope our method inspires future research at the intersection of computer graphics and generative AI, enabling safe deployment of these systems in the real world.

Acknowledgements. The authors acknowledge resources provided by Research Computing at Arizona State University. The authors also acknowledge technical access and support from ASU Enterprise Technology. This work was supported by NSF Robust Intelligence program grants #1750082 and #2132724. TG was supported by Microsoft's Accelerating Foundation Model Research (AFMR) program and UMBC's Strategic Award for Research Transitions (START). The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

1. Bai, J., et al.: Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond (2023). https://arxiv.org/abs/2308.12966
2. Bavishi, R., et al.: Introducing our multimodal models (2023). https://www.adept.ai/blog/fuyu-8b
3. Chatterjee, A., et al.: Getting it right: improving spatial consistency in text-to-image models (2024). https://arxiv.org/abs/2404.01197
4. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: attention-based semantic guidance for text-to-image diffusion models. ACM Trans. Graph. (TOG) **42**(4), 1–10 (2023)
5. Chen, J., et al.: Pixart-*alpha*: fast training of diffusion transformer for photorealistic text-to-image synthesis. In: The Twelfth International Conference on Learning Representations (2023)
6. Chen, M., et al.: Generative pretraining from pixels. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 1691–1703. PMLR (2020). http://proceedings.mlr.press/v119/chen20s.html
7. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance (2023)
8. Cho, J., Zala, A., Bansal, M.: Dall-eval: probing the reasoning skills and social biases of text-to-image generation models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3043–3054 (2023)
9. Dai, W., et al.: InstructBLIP: towards general-purpose vision-language models with instruction tuning. In: Thirty-Seventh Conference on Neural Information Processing Systems (2023). https://openreview.net/forum?id=vvoWPYqZJA

10. Dayma, B., et al.: Dall·e mini (2021). https://doi.org/10.5281/zenodo.5146400, https://github.com/borisdayma/dalle-mini
11. Feng, W., et al.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=PUIqjT4rzq7
12. Feng, W., et al.: LayoutGPT: compositional visual planning and generation with large language models. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
13. Fu, C., et al.: MME: a comprehensive evaluation benchmark for multimodal large language models. arXiv preprint **abs/2306.13394** (2023), https://arxiv.org/abs/2306.13394
14. Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: VQA-LOL: visual question answering under the lens of logic. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12366, pp. 379–396. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_23
15. Gokhale, T., et al.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)
16. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
17. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=_CDixzkzeyb
18. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: a comprehensive benchmark for open-world compositional text-to-image generation. In: Advances in Neural Information Processing Systems, vol. 36, pp. 78723–78747 (2023)
19. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: 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, 21–26 July 2017, pp. 1988–1997. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.215
20. Kamath, A., Hessel, J., Chang, K.W.: What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 9161–9175 (2023)
21. Li, Y., et al.: Gligen: open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22511–22521 (2023)
22. Li, Z., et al.: Super-clevr: a virtual benchmark to diagnose domain robustness in visual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14963–14973 (2023)
23. Lian, L., Li, B., Yala, A., Darrell, T.: LLM-grounded diffusion: enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint abs/2305.13655 (2023). https://arxiv.org/abs/2305.13655
24. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems, vol. 36 (2024)

26. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13677, pp. 423–439. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19790-1_26
27. Liu, Y., et al.: Mmbench: is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 (2023)
28. Meng, C., et al.: Sdedit: guided image synthesis and editing with stochastic differential equations. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, 25–29 April 2022. OpenReview.net (2022). https://openreview.net/forum?id=aBsCjcPu_tE
29. Minderer, M., et al.: Simple open-vocabulary object detection. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13670, pp. 728–755. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20080-9_42
30. Nichol, A.Q., et al.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17–23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 16784–16804. PMLR (2022). https://proceedings.mlr.press/v162/nichol22a.html
31. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings, pp. 1–11 (2023)
32. Patel, M., Kim, C., Cheng, S., Baral, C., Yang, Y.: Eclipse: a resource-efficient text-to-image prior for image generations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9069–9078 (2024)
33. Podell, D., et al.: SDXL: improving latent diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=di52zR8xgf
34. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021). http://proceedings.mlr.press/v139/radford21a.html
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents, **1**(2), 3. arXiv preprint arXiv:2204.06125 (2022)
36. Ramesh, A., et al.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8821–8831. PMLR (2021). http://proceedings.mlr.press/v139/ramesh21a.html
37. Razzhigaev, A., et al.: Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 286–295 (2023)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)

39. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems, vol. 35, pp. 36479–36494 (2022)
40. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, 5–10 December 2016, Barcelona, Spain, pp. 2226–2234 (2016). https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html
41. Sampat, S.K., Kumar, A., Yang, Y., Baral, C.: CLEVR_HYP: a challenge dataset and baselines for visual question answering with hypothetical actions over images. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3692–3709. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.289, https://aclanthology.org/2021.naacl-main.289
42. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, 7–12 December 2015, Montreal, Quebec, Canada, pp. 3483–3491 (2015). https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html
43. Stammer, W., Schramowski, P., Kersting, K.: Right for the right concept: revising neuro-symbolic concepts by interacting with their explanations. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, 19–25 June 2021, pp. 3619–3629. Computer Vision Foundation/IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00362, https://openaccess.thecvf.com/content/CVPR2021/html/Stammer_Right_for_the_Right_Concept_Revising_Neuro-Symbolic_Concepts_by_Interacting_CVPR_2021_paper.html
44. Team, G., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint abs/2312.11805 (2023). https://arxiv.org/abs/2312.11805
45. Wake, N., Kanehira, A., Sasabuchi, K., Takamatsu, J., Ikeuchi, K.: Gpt-4v (ision) for robotics: multimodal task planning from human demonstration. arXiv preprint arXiv:2311.12015 (2023)
46. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: unleashing the power of large-scale unlabeled data. arXiv preprint abs/2401.10891 (2024). https://arxiv.org/abs/2401.10891
47. Yang, L., et al.: Improving diffusion-based image synthesis with context prediction. In: Advances in Neural Information Processing Systems, vol. 36 (2024)
48. Yang, Z., et al.: The dawn of LMMS: preliminary explorations with gpt-4v (ision), **9**(1), 1. arXiv preprint arXiv:2309.17421 (2023)
49. Yang, Z., et al.: Reco: region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14246–14255 (2023)
50. Yi, K., et al.: CLEVRER: collision events for video representation and reasoning. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020. OpenReview.net (2020). https://openreview.net/forum?id=HkxYzANYDB

51. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023)
52. Zhang, R., et al.: Llama-adapter: efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
53. Zhang, T., Zhang, Y., Vineet, V., Joshi, N., Wang, X.: Controllable text-to-image generation with gpt-4. arXiv preprint arXiv:2305.18583 (2023)
54. Zhang, X., et al.: Realcompo: dynamic equilibrium between realism and compositionality improves text-to-image diffusion models. arXiv preprint arXiv:2402.12908 (2024)