SON: Enhancing Prompt Understanding of Diffusion Models with Large Language Models Guided Layouts

Weiyue Li wel019@ucsd.edu
Yi Li yil115@ucsd.edu
Xiaoyue Wang xiw027@ucsd.edu
Hao Zhang haozhang@ucsd.edu

Abstract

The recent development of text-to-image (T2I) models has unlocked numerous possibilities for content creation, particularly by offering inspiration to designers. However, current approaches often face challenges in accurately following prompts to generate images. These challenges include arranging non-overlapping objects in various spatial relationships and producing the correct number of desired objects, both of which are crucial for many design tasks. We introduce **S**patial-**O**verlap-**N**umeracy-1K (**SON-1K**), a comprehensive benchmark for text-to-image generation. This benchmark comprises 1,000 complex prompts spanning three subtasks: spatial relationships, numeracy counts, and complex natural prompts. Alongside the benchmark, we propose several evaluation metrics to assess compliance with the prompts comprehensively. We also propose a new approach, the **L**anguage **M**odel-Guided **D**iffusion++ (LMDpp), enhancing the performance of the novel two-stage Large Language Model (LLM)-grounded diffusion model pipeline (LMD). We report experimental results of previous major T2I models and our enhanced LMDpp, along with its baseline on SON-1K, and provide an analysis of our new metrics.

Website: https://weiyueli7.github.io/SON/
Code: https://github.com/weiyueli7/SON

1 Introduction

Numerous studies in the domain of compositional text-to-image (T2I) generation have concentrated on specific challenges, including attribute binding (Feng et al. 2023a; Chefer et al. 2023) and the depiction of spatial relations (Gokhale et al. 2022; Wu et al. 2023), each introducing separate benchmarks for method evaluation. Studies like (Huang et al. 2023; Feng et al. 2023b; Cho, Zala and Bansal 2023; Bakr et al. 2023) have developed comprehensive benchmarks for assessing open-world compositional T2I generation. Despite achieving high marks on these benchmarks, the latest advancements in T2I generation (Li et al. 2023; Podell et al. 2023; Wang et al. 2023; OpenAI 2023b) offer a bright outlook for the creation of diverse, high-quality images from naturally phrased prompts. However, these methods frequently encounter difficulties in accurately positioning a precise count of non-overlapping objects within correct spatial relationships, which is crucial for generating designers' sketches. See Appendix A.1 for more details.

To address these compositional T2I issues, we propose **S**patial-**O**verlap-**N**umeracy-1K (**SON-1K**) benchmark, consisting of three complex tasks: (1) **Spatial Reasoning**, with 400 prompts where each prompt consists of n distinct objects and $n-1$ spatial relationships, where $3 \leq n \leq 10$. (2) **Numerical Reasoning**, with 400 prompts in total, 200 of which have more than two categories of objects with varying numbers. (3) **Complex Natural Prompts**, with 200 prompts in total aimed at simulating human-like input to generate complex synthetic images. We also propose comprehensive metrics to assess the overall accuracy of the generated image in following the prompt input.

Researchers have started to use LLMs like ChatGPT (Ouyang et al. 2022) as tools for generating visual layouts using their reasoning capabilities (Wu et al. 2023; Feng et al. 2023b; Lian et al. 2023) and integrate these into existing region-controlled pipelines, showcasing state-of-the-art performance on various comprehensive compositional benchmarks. In addition to these 2-stage pipelines, we propose a new approach named **L**anguage **M**odel-Guided **D**iffusion++ (LMDpp), utilizing various proven prompting techniques to enhance the performance of Lian et al. (2023). Figure 1 showcases the example performance of our LMDpp method.

In summary, our main contributions are:

- We introduce the SON-1K benchmark, offering more complex challenges in spatial reasoning, numeracy, and natural language processing than current benchmarks, along with comprehensive metrics that account for object overlapping.
- We enhance a two-stage LLM-grounded diffusion model pipeline through strategic prompting and showcase the improvements of our LMDpp pipeline in object spatial relationships, spacing, and numeracy tasks.
- Through benchmarking various state-of-the-art T2I pipelines, we conclude that using LLMs as tools for 2-stage T2I pipelines has a promising future for generating images that could inspire designers in their productions.

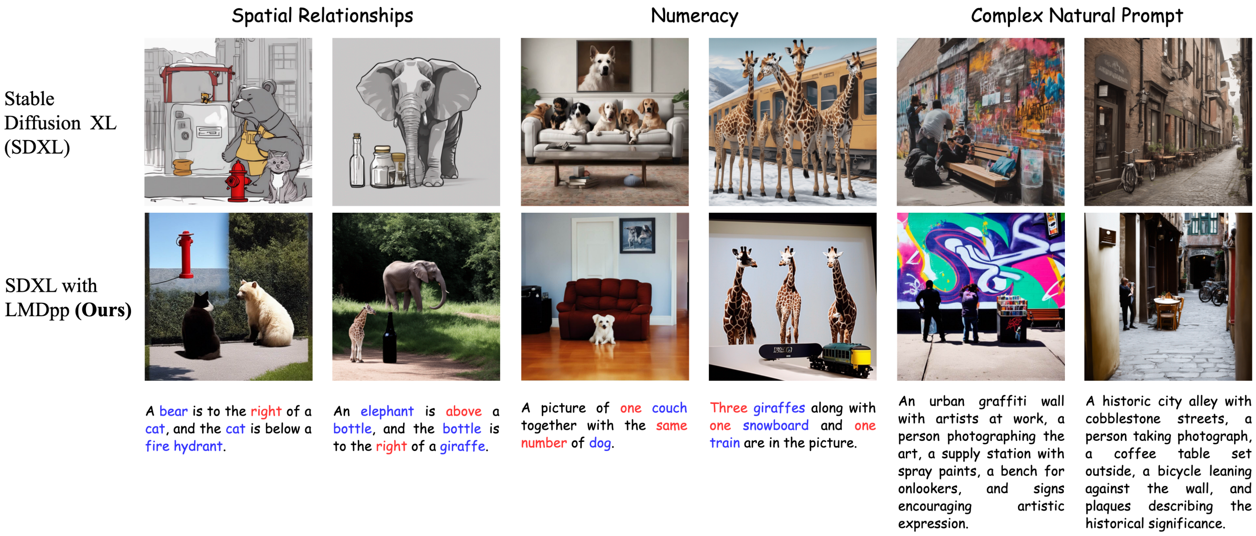

Spatial Relationships
Numeracy
Complex Natural Prompt
Stable Diffusion XL (SDXL)
SDXL with LMDpp (Ours)
A bear is to the right of a cat, and the cat is below a fire hydrant.
An elephant is above a bottle, and the bottle is to the right of a giraffe.
A picture of one couch together with the same number of dog.
Three giraffes along with one snowboard and one train are in the picture.
An urban graffiti wall with artists at work, a person photographing the art, a supply station with spray paints, a bench for onlookers, and signs encouraging artistic expression.
A historic city alley with cobblestone streets, a person taking photograph, a coffee table set outside, a bicycle leaning against the wall, and plaques describing the historical significance.

Figure 1: Our proposed LMDpp method attains significant performance enhancement over the baseline Stable Diffusion XL (Podell et al. 2023) on all tasks of our SON-1K benchmark.

2 Related Work

Text-to-image (T2I) generation is to create an image that matches a given textual description. Efforts to enhance the quality of the generated image include various network architectures and loss functions (Creswell et al. 2018). More recently, diffusion models have attracted significant attention for achieving superior results in this generation task. DALL-E (Ramesh et al. 2021) and DALL-E-3 (OpenAI 2023b), with its multimodal latent space, demonstrates exceptional performance in image generation, surpassing previous models. The latest open-source state-of-the-art Stable Diffusion Model (SDXL) (Podell et al. 2023) is capable of producing high-quality images. However, these models encounter challenges in precisely retaining positional and numerical information from the original description. Works like TokenCompose (Wang et al. 2023) improve multi-category instance composition by introducing the token-wise consistency terms between the image content and object segmentation, and Attend-and-Excite pipeline (Chefer et al. 2023) guides the model to attend all subject tokens in prompt, but the spatial and numerical problems still remain.

Large language models (LLMs) (Brown et al. 2020) has rapidly evolved in recent years. Leveraging the robust capabilities of LLMs, (Cai et al. 2023) employs LLMs as Tool Makers for problem-solving, and (Bai et al. 2023; Liu et al. 2023; Hu et al. 2023) utilize them as visual question-answering tools. The exploration of Chain-of-Thoughts reasoning (Kojima et al. 2022; Wei et al. 2022; Wang et al. 2022; Huang et al. 2022), prompt-engineering (Gao, Fisch and Chen 2021; Liu et al. 2023; Ouyang et al. 2022; Wei et al. 2021; Sanh et al. 2022), and query generation (Wang et al. 2024) techniques enable LLMs like ChatGPT serve as good reasoner to undertake annotation and generation tasks (Zhang, Wang and Shang 2023; Ding et al. 2022; Wang et al. 2024).

Image layout generation is an important task for helping various design tasks like indoor design (Ritchie, Wang and Lin 2019; Wang et al. 2019) or document layout capture (Zheng et al. 2019), and the visual layouts, which reflect the compositions of a visual space (Luo et al. 2020; Yang et al. 2021; Wang et al. 2022; Ma et al. 2023), have been widely studied. Other layout generation models (Jyothi et al. 2019; Li et al. 2019; Gupta et al. 2021; Yang et al. 2021; Kong et al. 2022) can be combined with region-controlled image generation methods (Yang et al. 2023; Li et al. 2023) to improve image compositionality (Wu et al. 2023). However, these models are restricted to discrete categories or have limited reasoning skills for complicated text conditions. Recently, LayoutGPT (Feng et al. 2023b) and LLM-Grounded Diffusion (LMD) (Lian et al. 2023) have both incorporated Large Language Models (LLMs) as the reasoner and have introduced novel region-controlled methods to enhance the diffusion model by grounding input information through input regularization and providing layout structures to the diffusion models.

The compositional image generation benchmarks offer T2I pipelines for validating compositional issues like missing objects, incorrect attributes, and incorrect spatial relations. The available benchmark VISOR (Gokhale et al. 2022), HRS-Bench (Bakr et al. 2023), and T2I-CompBench (Huang et al. 2023) offers datasets for evaluating spatial relationships in text-to-image generation. However, those benchmarks typically feature one or two objects and a single spatial relationship between them, which does not simulate the complex scene with multiple objects and their relations. Additionally, the benchmark NSR-1k, introduced in LayoutGPT, and PaintSkills (Cho, Zala and Bansal 2023), are datasets with spatial and counting annotations to assess layout generation quality. However, those datasets also only include fewer than three different object categories.

3 SON-1K

Current state-of-the-art diffusion models and pipelines are capable of generating realistic and diverse images from text prompts. However, they often struggle with prompts that include complex spatial relationships, varying numbers of desired objects, and specific instructions such as preventing object overlap. While existing benchmarks do include spatial and numeracy tasks, these compositional tasks are relatively simple, typically involving only two distinct object categories. This simplicity enables current models to perform well, but it does not accurately represent more complex real-life scenarios where users may want to create images with more than two object categories, each requiring specific spatial relationships and numeracy values.

To bridge the gap between current compositional benchmarking practices and real-life use cases, we have developed the **S**patial-**O**verlap-**N**umeracy-1K (**SON-1K**), which is designed to assess the performance of models on comprehensive text-to-image generation tasks. It comprises three datasets focused on spatial (400 prompts), numeracy (400 prompts), and complex natural prompt (200 prompts) tasks. Each task features complex prompts that involve more than two distinct object categories. These prompts were created by either expanding upon existing benchmarks (Gokhale et al. 2022; Feng et al. 2023b; Lin et al. 2015) to include more object categories and their corresponding relationships or by utilizing

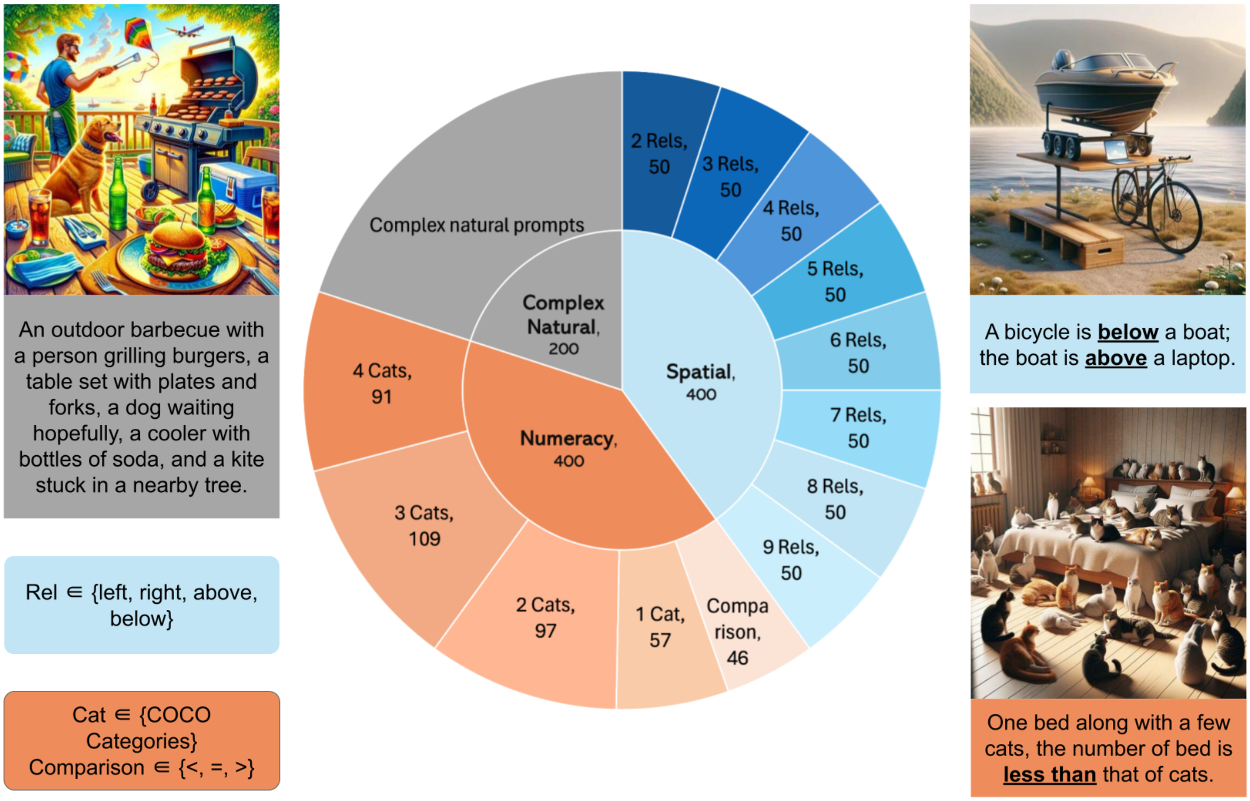


Figure 2: Overview of our SON-1k benchmark.

ChatGPT (OpenAI 2023a), followed by human selection for refinement. We make the source code available to generate these data points, facilitating easy scaling. All possible object categories in our benchmarks are from the MS-COCO (Lin et al. 2015) categories. The distribution and visual representation of SON-1K is detailed in Figure 2. Additionally, we have compiled Table 1 to showcase comparisons between our benchmark and the existing works.

3.1 Spatial Relationships

Being able to strictly follow the spatial relationships between objects is a critical aspect of the quality of synthetic images generated by T2I pipelines. The spatial dataset is designed to assess the model's performance in generating objects that adhere to the spatial relationships specified in the prompt. While previous works focused on evaluating two objects with one spatial relationship within a single prompt, we expanded upon VISOR by including more objects (up to 10) and additional spatial relationships. This task comprises 400 data points. For any two objects, the options for spatial relationships are {“to the left of”, “to the right of”, “above”, “below”}. The number of spatial relationships in a prompt ranges from 2 to 9, corresponding to 3 to 10 objects with 50 prompts for each configuration. Thus,

Table 1: Comparison of compositional text-to-image benchmarks.

Benchmark	**Prompts number and tasks**	**Object categories**	**Object categories/prompt**
VISOR (Gokhale et al. 2022)	25,280 rel (spatial)	80 (MS-COCO)	2
HRS-Bench (Bakr et al. 2023)	2,000 rel (spatial)	700	2 (**NEED TO CHECK**)
T2I-CompBench (Huang et al. 2023)	2,000 rel (spatial), 1,000 complex	80 (MS-COCO)	2
PaintSkills (Cho, Zala and Bansal 2023)	2700 rel (spatial), 2160 numeracy	15 (MS-COCO)	1-2 (spatial) 1 (numeracy)
NSR-1K (Feng et al. 2023b)	283 rel (spatial), 762 numeracy	80 (MS-COCO)	1-2 (spatial) 1-2 (numeracy)
SON-1K	400 rel (spatial), 400 numeracy, 200 complex	80 (MS-COCO)	3-10 (spatial) 1-4 (numeracy)

for a prompt with n objects, it will have $n-1$ spatial relationships Rel such that:

$$Rel(O_i, O_{i+1}) \in \{\text{“to the left of”}, \text{“to the right of”}, \text{“above”}, \text{“below”}\},$$
$$O_i, O_{i+1} \in \{\text{MS-COCO Classes}\} \forall i \in [1, n-1].$$

For example, for $n = 4$, a prompt might be: “A realistic scene with 4 objects ([‘spoon’, ‘knife’, ‘microwave’, ‘apple’]): the spoon is *below* the knife; the knife is *to the right of* the microwave; the microwave is *above* the apple.”

3.2 Numeracy

In addition to incorrect spatial relationships during the text-to-image (T2I) generation process, generating an inaccurate number of objects represents another significant challenge in T2I compositional tasks. The numerical reasoning dataset, which has 400 data points, is crafted to evaluate a model’s ability to generate objects that match the precise quantities specified in the prompt. Previous research has primarily concentrated on prompts involving only one category of objects in varying numbers (referred to as “one category”), two categories of distinct objects in varying numbers (“two categories”), and prompts including two categories of distinct objects but specifying the number for only one type, with the quantity of the other type inferred through comparative terms like “fewer than,” “an equal number of,” and “more than” (”comparison”). To introduce more intricate numerical reasoning challenges, we have expanded the scope of NSR-1K to include prompts with three and four categories of distinct objects from MS-COCO, each in varying quantities.

3.3 Complex Natural Prompts

The complex natural prompts dataset is curated to simulate the kinds of descriptions individuals typically use when they aim to generate an image of a specific scene. Unlike structured tasks, these natural prompts do not adhere to a predetermined format. To achieve a flow in the descriptions that closely resembles human communication, we created 400 data points for this task using ChatGPT-4, with potential objects drawn from the MS-COCO classes. Following the initial generation of prompts, we conducted a human evaluation to assess the quality of these prompts. To maintain the integrity of our dataset and align with our goal of creating complex natural tasks, we excluded any prompts containing fewer than three objects. This filtering process has resulted in 200 high-quality complex natural prompts that comprise our final dataset.

4 Evaluation Metrics

Intersection Over Union (IOU), also known as the Jaccard index, is a widely used evaluation metric in object detection benchmarks. It quantifies the extent to which the predicted bounding box aligns with the ground truth bounding box. In our work, we utilize IOU scores for two distinct purposes. The first is to determine the overlap rate between two bounding boxes. The second is to assess the alignment between the bounding boxes in the layout generated by LMD and the bounding boxes in the final images detected by YOLO v8 (Jocher et al. 2022). Given two bounding boxes, A and B, the IOU score is computed using the following equations (Rezatofighi et al. 2019):

$$IOU(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Given the centroids of bounding boxes (the coordinates of top-left and bottom-right are [0,0] and [width, height] respectively) in the final images produced by YOLO v8 (x_A, y_A) and (x_B, y_B), we move beyond simple comparisons of x or y coordinates for assessing the spatial relationship between two bounding boxes. Instead, we incorporate the Euclidean distance d along with sine and cosine values to achieve a more precise measurement of spatial relationships. For instance, if the difference between y_A and y_B is significantly larger than that between x_A and x_B, our metric prioritizes the above-below relationship over the left-right relationship. Furthermore, we include the IOU score from equation (1) to ensure that the overlap area between two bounding boxes is not excessively large. If the overlap area between two bounding boxes is substantial, then the spatial relationship becomes less meaningful. The detailed equation for our metric is defined as follows (Feng et al. 2023b; Huang et al. 2023):

$$Rel(A,B) = \begin{cases} B \text{ above } A & \text{if } \frac{y_B - y_A}{d} \geq sin(\pi/4) \text{ and } IOU(A,B) < 0.1 \\ B \text{ below } A & \text{if } \frac{y_B - y_A}{d} \leq sin(-\pi/4) \text{ and } IOU(A,B) < 0.1 \\ B \text{ on the left of } A & \text{if } \frac{x_B - x_A}{d} < cos(3\pi/4) \text{ and } IOU(A,B) < 0.1 \\ B \text{ on the right of } A & \text{if } \frac{x_B - x_A}{d} > cos(\pi/4) \text{ and } IOU(A,B) < 0.1 \\ \text{None} & \text{if } IOU(A,B) \geq 0.1 \end{cases} \tag{2}$$

Given N spatial relationships, we propose assessing the spatial accuracy rate by comparing the spatial relationship between two predicted bounding boxes in the final images produced by YOLO v8, as calculated using Equation (2), with their respective ground truth spatial relationships. The detailed equation of our metric is defined as follows:

$$\text{Spa-Acc} = \frac{1}{N}\sum(Rel(a,b)_{PRED} == Rel(a,b)_{GT}) \tag{3}$$

To better assess the compliance between generated images and their corresponding prompts, we have designed several new metrics to evaluate the images from three different perspectives: object spatial accuracy, object overlap rate, and object numeracy accuracy.

We introduce the equation Spa-Lap, which incorporates spatial accuracy and overlap to evaluate the predicted bounding boxes in the final images produced by YOLO v8. For each score, we assign a weight between 0 and 1. If the generation task prioritizes spatial accuracy over the overlap rate, then we will apply a higher weight to spatial accuracy. In our case, we use $w = 0.5$, and the detailed equation of our metric is defined as follows:

$$\textbf{Spa-Lap} = (w)\cdot(\text{Spa-Acc}) + (1-w)\cdot(1-\%\ \text{Overlap}) \tag{4}$$

Where the overlap rate is calculated as:

$$\%\ \text{Overlap} = \frac{\sum_{i=0}^{N} \text{IOU (A, B)}_i}{N} \tag{5}$$

We denote the set of n (m) object classes in the ground truth (prediction) as $C_{GT} = c_1, c_2, ..., c_n$ ($C_{Pred} = c'_1, c'_2, ..., c'_m$), where $x_{c_1}, x_{c_2}, ..., x_{c_n}$ ($x'_{c'_1}, x'_{c'_2}, ..., x'_{c'_m}$) represent the number of objects for each category accordingly. We define the Error Miss Ratio (EMR) as the percentage of total detected objects that differ from the ground truth annotations:

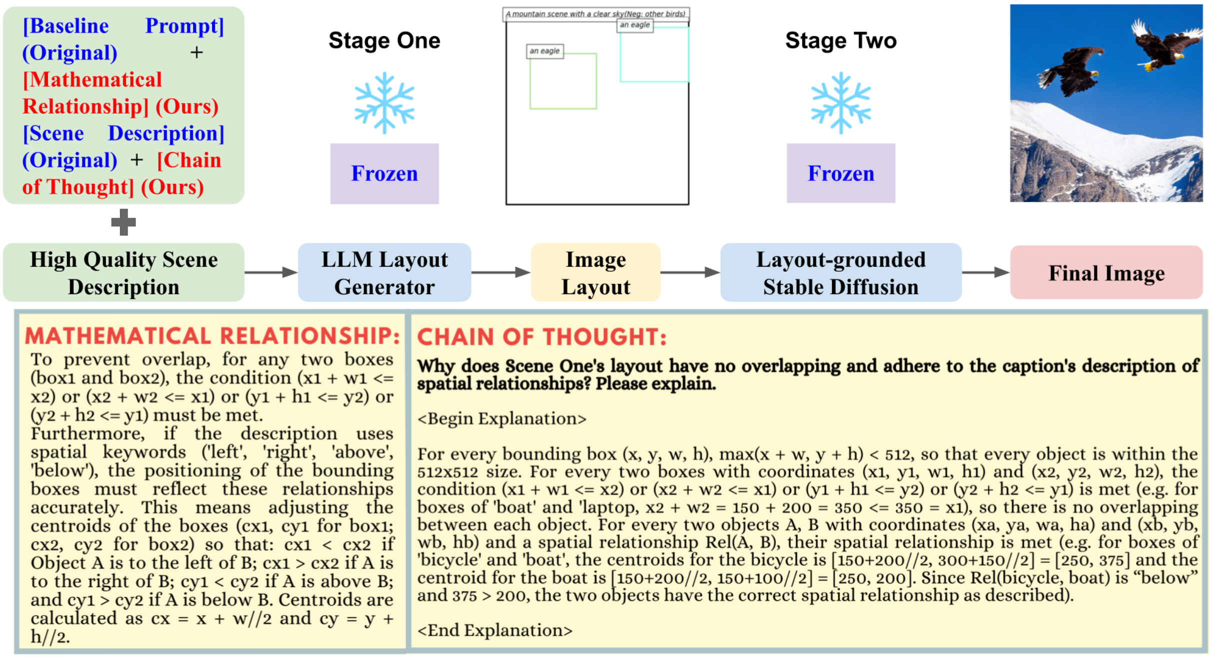
$$\text{EMR} = \frac{|\sum x'_{c'_k} - \sum x_{c_k}|}{\sum x_{c_k}} \tag{6}$$

We also introduce the equation Num-Lap, which incorporates numeracy accuracy and overlap to evaluate the predicted bounding boxes in the final images produced by YOLO v8. The metric we use to evaluate numeracy accuracy is the EMR. The detailed equation of our metric is defined as follows:

$$\textbf{Num-Lap} = (1 - \frac{1}{N}\sum_{i=0}^{N}(\text{EMR})_i)\cdot(1-\%\ \text{Overlap}) \tag{7}$$

Then, we combine the equations Spa-Lap and Num-Lap into a comprehensive metric called SON (**S**patial-**O**verlap-**N**umeracy). For the spatial task, we take the average of the results from Spa-Lap and Num-Lap. For numeracy and complex natural tasks, we only measure Num-Lap since the main focus of numeracy and complex natural tasks is not spatial accuracy. The detailed equation of SON is defined as follows:

$$\textbf{SON (Spatial-Overlap-Numeracy)} = \begin{cases} \frac{1}{2}\cdot\text{Spa-Lap} + \frac{1}{2}\cdot\text{Num-Lap} & \text{if task is spatial} \\ \text{Num-Lap} & \text{else} \end{cases} \tag{8}$$


[Baseline Prompt] (Original) + [Mathematical Relationship] (Ours)
[Scene Description] (Original) + [Chain of Thought] (Ours)
Stage One
Frozen
A mountain scene with a clear sky(Neg: other birds)
an eagle
an eagle
Stage Two
Frozen
High Quality Scene Description
LLM Layout Generator
Image Layout
Layout-grounded Stable Diffusion
Final Image

MATHEMATICAL RELATIONSHIP:

To prevent overlap, for any two boxes (box1 and box2), the condition (x1 + w1 <= x2) or (x2 + w2 <= x1) or (y1 + h1 <= y2) or (y2 + h2 <= y1) must be met.
Furthermore, if the description uses spatial keywords ('left', 'right', 'above', 'below'), the positioning of the bounding boxes must reflect these relationships accurately. This means adjusting the centroids of the boxes (cx1, cy1 for box1; cx2, cy2 for box2) so that: cx1 < cx2 if Object A is to the left of B; cx1 > cx2 if A is to the right of B; cy1 < cy2 if A is above B; and cy1 > cy2 if A is below B. Centroids are calculated as cx = x + w//2 and cy = y + h//2.

CHAIN OF THOUGHT:

Why does Scene One's layout have no overlapping and adhere to the caption's description of spatial relationships? Please explain.

<Begin Explanation>

For every bounding box (x, y, w, h), max(x + w, y + h) < 512, so that every object is within the 512x512 size. For every two boxes with coordinates (x1, y1, w1, h1) and (x2, y2, w2, h2), the condition (x1 + w1 <= x2) or (x2 + w2 <= x1) or (y1 + h1 <= y2) or (y2 + h2 <= y1) is met (e.g. for boxes of 'boat' and 'laptop, x2 + w2 = 150 + 200 = 350 <= 350 = x1), so there is no overlapping between each object. For every two objects A, B with coordinates (xa, ya, wa, ha) and (xb, yb, wb, hb) and a spatial relationship Rel(A, B), their spatial relationship is met (e.g. for boxes of 'bicycle' and 'boat', the centroids for the bicycle is [150+200//2, 300+150//2] = [250, 375] and the centroid for the boat is [150+200//2, 150+100//2] = [250, 200]. Since Rel(bicycle, boat) is “below” and 375 > 200, the two objects have the correct spatial relationship as described).

<End Explanation>

Figure 3: Pipeline of our LMDpp method.

5 Methods

We adopt the LLM-Grounded Diffusion pipeline (LMD) (Lian et al. 2023) as the backbone for our LMDpp pipeline, shown in Figure 3. The overall pipeline structure of LMD consists of two stages. In the first stage, LMD takes the user prompt as input and converts it into a template with instructions and in-context examples. The LLM is then prompted to complete the scene descriptions, following the style of the in-context examples. Finally, the LLM's output is parsed into a set of object-bounding boxes and a background caption. In the second stage, a stable diffusion model generates the final images under the guidance of the layout-grounded controller. Both the LLM and the Diffusion Model are frozen, which makes LMD adaptable to different LLMs and Diffusion Models without needing to delve into their training objectives.

However, as we apply the LMD pipeline to our SON-1K benchmarks, we find that there is still considerable room for improvement in spatial and numeracy accuracy and in reducing the overlap rate. For instance, as the number of objects increases, the positions of the objects and the spatial relationships among them are not satisfactory. At the same time, the overlap rate also significantly increases. As a result, we decide to develop a solution for better image generation, meaning more precise spatial and numeracy accuracy and a lower overlap rate for scenes with a varying number of objects, different object types, and spatial relationships.

The essential part of our strategy is to utilize proven prompting techniques to provide better guidance so that the LLM can adhere to the instructions. We apply two techniques for prompt engineering: (1) inserting mathematical relationships to enforce the position and spatial relationship among objects and (2) incorporating the chain of thought to improve the LLM's reasoning on the mathematical relationship between bounding boxes' coordinates.

5.1 Mathematical Relationship

We include eight mathematical relationships to teach the LLM how to determine the spatial distance between two bounding boxes and whether there is overlap between them. More specifically, we have four mathematical relationships, as shown in Figure 3, to compare the x and y coordinates of objects' centroids to help the LLM determine their spatial relationships. For the overlapping part, we also incorporate four mathematical relations as conditions to help the LLM understand under what circumstances two bounding boxes will contain no overlap.

5.2 Chain of Thought

We redesign three in-context examples utilizing the chain of thought technique. For each example, we adopt the same scene caption style as the one in our spatial benchmarks to help the LLM perform more efficient few-shot learning (Brown et al. 2020). Furthermore, for each example, we provide a step-by-step explanation to help the LLM understand why the coordinates for each bounding box adhere to the scene description. Incorporating step-by-step explanations for each mathematical relation will increase the LLM's capabilities in processing those equations, leading to images with better spatial accuracy and less overlap. For complete in-context examples, see Appendix A.2

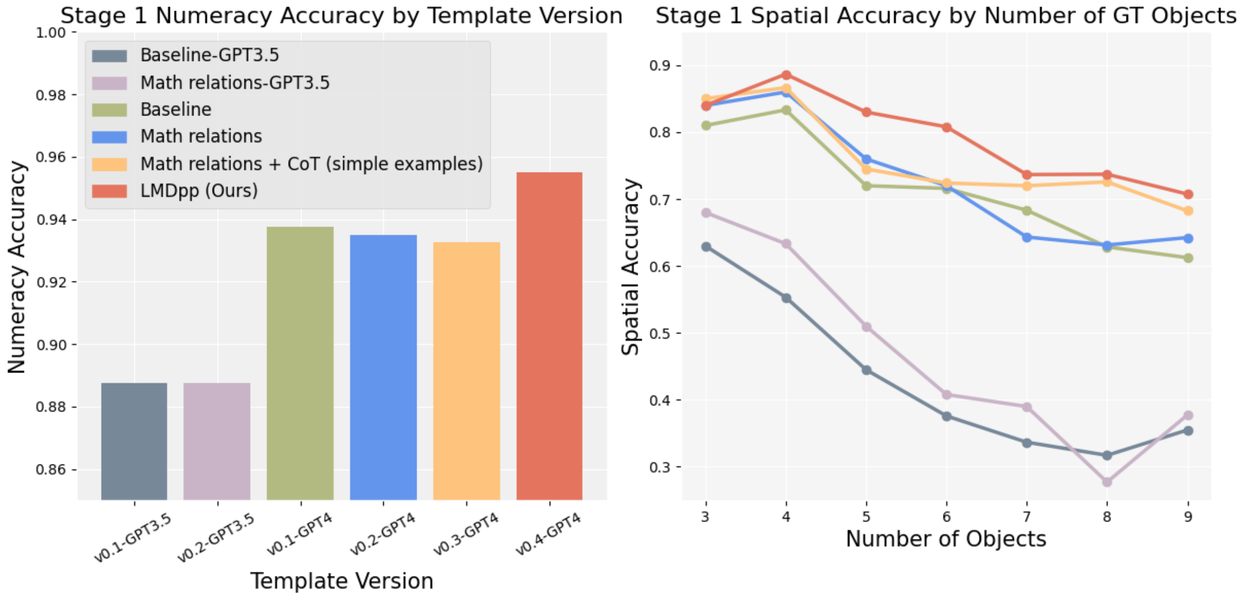


Figure 4: Comparison of numeracy & spatial accuracy among different templates of prompt.

6 Results

6.1 Layout Generation on SON-1K

We compare our LMDpp against several other pipelines in both numeracy and spatial accuracy: (1) **Baseline-GPT3.5**, which utilizes the LMD framework with GPT3.5 as the LLM; (2) **Math Relations-GPT3.5**, Baseline-GPT3.5 with incorporated mathematical relations in prompt; (3) **Baseline (with GPT4)**; (4) **Math Relations (with GPT4)**; (5) **Math Relations + CoT (simple examples) (with GPT4)**, which incorporates both mathematical relationships and chain of thought into the prompts along with simple few-shot examples.

Based on Figure 4 (left), our model exceeds all other pipelines in numeracy accuracy, as LMDpp incorporates both mathematical relations and chain of thought in the prompt, along with well-designed few-shot examples.

Based on Figure 4 (right), our model surpasses all other pipelines in spatial accuracy with 4-9 objects present but falls short in scenes with 3 objects. We conjecture the reason may be that a scene with 3 objects is not complex enough to fully showcase our model's capabilities. The performance between our model and others is quite close. As the number of present objects increases, the overall trend of spatial accuracy decreases, but our model remains the best.

Table 2: Evaluation results for Spatial Tasks. % Overlap (S1) indicates the overlap rate for the layout image generated by the LLM in stage one. Spa-Acc (S1) indicates the spatial accuracy for stage one. SON indicates the values measured by our metric SON.

Model	LLM	% Overlap (S1)	Spa-Acc (S1)	SON
LMD	GPT4	0.023	0.695	0.531
LMDpp (ours)	GPT4	**0.016 (**↓ 30.4%**)**	**0.772 (**↑ 11.1%**)**	**0.540**
LMD	GPT3.5T	0.105 (↑ 356.5%)	0.416 (↓ 40.1%)	0.522
LMDpp (ours)	GPT3.5T	0.071 (↑ 208.7%)	0.445 (↓ 36.0%)	0.538

Table 3: Evaluation results for Numeracy Tasks. Num-Recall (S1) indicates the numeracy recall score for stage one. Num-Acc (S1) indicates the numeracy accuracy for stage one.

Model	LLM	Num-Recall (S1)	Num-Acc (S1)	% Overlap (S1)	SON
LMD	GPT4	0.983	0.938	0.071	0.060
LMDpp (ours)	GPT4	**0.991 (**↑ 0.8%**)**	**0.955 (**↑ 1.8%**)**	**0.014 (**↓ 80.3%**)**	0.049
LMD	GPT3.5T	0.965 (↓ 1.83%)	0.888 (↓ 5.3%)	0.056 (↓ 21.1%)	0.037
LMDpp (ours)	GPT3.5T	0.967 (↓ 1.6%)	0.888 (↓ 5.3%)	0.158 (↑ 122.5%)	**0.064**

Table 4: Evaluation results for Complex Prompts. EMR (S2) indicates the EMR score of the final image generated by the diffusion model in stage two.

Model	LLM	% Overlap (S1)	EMR (S2)	SON
LMD	GPT4	0.067	0.624	0.369
LMDpp (ours)	GPT4	**0.036** (↓ 46.3%)	**0.537** (↓ 13.9%)	**0.454**
LMD	GPT3.5T	0.119 (↑ 77.6%)	0.646 (↑ 3.5%)	0.344
LMDpp (ours)	GPT3.5T	0.150 (↑ 123.9%)	0.681 (↑ 9.1%)	0.311

6.2 Image Generation on SON-1K

In Tables 2, 3, and 4, we compare our LMDpp pipeline against the baseline LMD pipeline in terms of overlap rate, spatial accuracy, numeracy recall & accuracy for stage 2, and EMR score and SON for stage 2. In addition to GPT4, we also include the results of using GPT3.5-turbo as part of our ablation study.

Spatial Task: Based on Table 2, our model LMDpp outperforms the baseline in % Overlap (S1), Spa-Acc (S1), and SON score. This trend is maintained as we switch our LLM from GPT4 to GPT3.5-turbo.

Numeracy Task: Based on Table 3, our model LMDpp outperforms the baseline in Num-Recall (S1), Spa-Acc (S1), and % Overlap (S1). This trend is maintained as we switch our LLM from GPT4 to GPT3.5-turbo, except for the % Overlap (S1). We conjecture that this is due to the limited token size of GPT3.5-turbo. Our refined prompt for LMDpp is significantly longer than the one used for LMD. If GPT3.5-turbo cannot take the full prompt, then the incomplete refined prompt will be confusing for the LLM. Also, we observe that LMDpp with GPT4 achieves a lower SON score than LMDpp with GPT3.5-turbo. We conjecture that it could be that GPT-3.5 has been unintentionally optimized for certain types of image-related tasks or prompts during training.

Complex Natural Task: Based on Table 4, our model LMDpp outperforms the baseline in % Overlap (S1), EMR (S2), and SON score. This trend is completely flipped as we switch our LLM from GPT4 to GPT3.5-turbo. Similarly, we conjecture that the limited token size of GPT3.5-turbo could be the essential reason, as we discussed earlier.

Table 5: Comparison between LMDpp with other T2I pipelines. Spa-Acc (S2) indicates the spatial accuracy for stage two. % Overlap (S2) indicates the overlap rate for stage two.

Model	Spa-Acc (S2)	EMR (S2)	% Overlap (S2)	SON
LMD	0.145	0.492	0.018	0.531
LMDpp (ours)	0.170 (↑ 17.2%)	0.485 (↓ 1.4%)	0.020 (↑ 11.1%)	0.540
GLIGEN	0.059 (↓ 59.3%)	0.580 (↑ 17.9%)	0.030 (↑ 66.7%)	0.455
SDXL	0.041 (↓ 71.7%)	0.597 (↑ 21.3%)	0.015 (↓ 16.6%)	0.461
TokenCompose	0.020 (↓ 86.2%)	0.588 (↑ 19.5%)	0.061 (↑ 238.8%)	0.433
DALL-E-3	**0.182** (↑ 25.5%)	**0.300** (↓ 39.0%)	**0.013** (↓ 27.8%)	**0.638**

6.3 Pipeline Comparison

In Table 5, we compare our LMDpp pipeline against several others in Spa-Acc (S2), EMR (S2), % Overlap (S2), and SON score: (1) **Original LMD Pipeline**; (2) **GLIGEN**; (3) **SDXL**; (4) **TokenCompose**; (5) **DALL-E-3**, where SDXL, TokenCompose, and DELL-E-3 directly take the scene prompt and generate the image.

Based on the results, DALL-E-3 outperforms all other pipelines, including LMD & LMDpp with layout image support, in Spa-Acc (S2), EMR (S2), and % Overlap (S2). This is expected because DALL-E-3 benefits significantly from a vast and diverse training dataset curated by OpenAI, which all open-source diffusion models have no access to. Besides that, LMDpp does outperform all open-source pipelines in Spa-Acc (S2), EMR (S2), and SON scores. However, it falls short in % Overlap (S2) compared to the SDXL framework. We conjecture that one possible reason is that our refined prompt is long, which could sometimes make the LLM hallucinate. Also, the layout-grounded controller could confuse stable diffusion models during the image generation process compared with directly generating images based on the text prompt.

7 Discussion and Conclusion

We introduce SON-1K, a comprehensive compositional text-to-image benchmark focusing on spatial relationships (400), numerical reasoning (400), and complex prompts (200), along with new metrics that consider object overlap in evaluation. These aim to address issues in compositional T2I generation and inspire designers in production. We provide the source code to facilitate the easy scaling of the dataset size, enabling a more comprehensive comparison between existing models. Additionally, we introduce a new method, LMDpp, which enhances the performance of the two-stage LMD pipeline through prompting techniques. Our study demonstrates that this enhanced two-stage technique surpasses all existing open-source T2I pipelines, while DALL-E-3 exhibits the best performance. This sheds light on future improvements, focusing on developing novel techniques for region control in diffusion models to generate more realistic and high-quality images, bringing us one step closer to achieving DALL-E-3's state-of-the-art performance.

References

Archicgi. 2019. “Floor Plan 3D Rendering Vs 2D Floor Plans – What Is Better?.” 06. [Link]

Bai, Jinze, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. “Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond.”

Bakr, Eslam Mohamed, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. 2023. “HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models.”

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell et al. 2020. “Language models are few-shot learners.” *Advances in neural information processing systems* 33: 1877–1901

Cai, Tianle, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2023. “Large language models as tool makers.” *arXiv preprint arXiv:2305.17126*

Chefer, Hila, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. “Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models.” *ACM Transactions on Graphics (TOG)* 42 (4): 1–10

Cho, Jaemin, Abhay Zala, and Mohit Bansal. 2023. “DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models.” In *ICCV*.

Creswell, Antonia, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. 2018. “Generative adversarial networks: An overview.” *IEEE signal processing magazine* 35 (1): 53–65

Ding, Bosheng, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. “Is GPT-3 a Good Data Annotator?” *arXiv preprint arXiv:2212.10450*

Etsy. 2024. “First Words Alphabet Flash Cards Baby Toddler Children Educational Learning SEN - Etsy.” [Link]

Feng, Weixi, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023a. “Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis.” In *ICLR*.

Feng, Weixi, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023b. “LayoutGPT: Compositional Visual Planning and Generation with Large Language Models.” *arXiv preprint arXiv:2305.15393*

Gao, Tianyu, Adam Fisch, and Danqi Chen. 2021. “Making Pre-trained Language Models Better Few-shot Learners.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Gokhale, Tejas, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. 2022. “Benchmarking Spatial Relationships in Text-to-Image Generation.” *arXiv preprint arXiv:2212.10015*

Gupta, Kamal, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2021. “Layouttransformer: Layout generation and completion with self-attention.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Hu, Wenbo, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2023. “BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions.”

Huang, Kaiyi, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. “T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation.” *arXiv preprint arXiv: 2307.06350*

Huang, Wenlong, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents.” In *International Conference on Machine Learning*. PMLR

Jocher, Glenn, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, □□□(Zeng Yifu), Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. 2022. “ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation.” November. [Link]

Jyothi, Akash Abdu, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. 2019. “Layoutvae: Stochastic scene layout generation from a label set.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Kojima, Takeshi, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. “Large Language Models are Zero-Shot Reasoners.” In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. [Link]

Kong, Xiang, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. 2022. “BLT: bidirectional layout transformer for controllable layout generation.” In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer

Li, Jianan, Tingfa Xu, Jianming Zhang, Aaron Hertzmann, and Jimei Yang. 2019. “LayoutGAN: Generating Graphic Layouts with Wireframe Discriminator.” In *International Conference on Learning Representations*. [Link]

Li, Yuheng, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. “GLIGEN: Open-Set Grounded Text-to-Image Generation.” *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

Lian, Long, Boyi Li, Adam Yala, and Trevor Darrell. 2023. “LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models.” *arXiv preprint arXiv:2305.13655*

Lin, Tsung-Yi, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. “Microsoft COCO: Common Objects in Context.”

Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. “Visual Instruction Tuning.”

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.” *ACM Computing Surveys* 55 (9): 1–35

Luo, Andrew, Zhoutong Zhang, Jiajun Wu, and Joshua B Tenenbaum. 2020. “End-to-end optimization of scene layout.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ma, Wan-Duo Kurt, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. 2023. “Directed Diffusion: Direct Control of Object Placement through Attention Guidance.” *arXiv preprint arXiv:2302.13153*

OpenAI. 2023a. https://openai.com/blog/chatgpt/

OpenAI. 2023b. “DALL-E 3.” https://openai.com/dall-e-3

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray et al. 2022. “Training language models to follow instructions with human feedback.” *Advances in Neural Information Processing Systems* 35: 27730–27744

Podell, Dustin, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. “Sdxl: Improving latent diffusion models for high-resolution image synthesis.” *arXiv preprint arXiv:2307.01952*

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. “Zero-shot text-to-image generation.” In *International conference on machine learning*. Pmlr

Rezatofighi, Hamid, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. “Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ritchie, Daniel, Kai Wang, and Yu-an Lin. 2019. “Fast and flexible indoor scene synthesis via deep convolutional generative models.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Sanh, Victor, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja et al. 2022. “Multitask Prompted Training Enables Zero-Shot Task Generalization.” In *ICLR 2022-Tenth International Conference on Learning Representations*.

Wang, Bo, Tao Wu, Minfeng Zhu, and Peng Du. 2022. “Interactive Image Synthesis with Panoptic Layout Generation.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, Jianyou Andre, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2024. “Scientific document retrieval using multi-level aspect-based queries.” *Advances in Neural Information Processing Systems* 36

Wang, Jianyou, Kaicheng Wang, Xiaoyue Wang, Weili Cao, Ramamohan Paturi, and

Leon Bergen. 2024. "IR2: Information Regularization for Information Retrieval."

Wang, Kai, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. 2019. "Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks." *ACM Transactions on Graphics (TOG)* 38 (4): 1–15

Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. "Self-consistency improves chain of thought reasoning in language models." *arXiv preprint arXiv:2203.11171*

Wang, Zirui, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. 2023. "Token-Compose: Grounding Diffusion with Token-level Supervision."

Wei, Jason, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. "Finetuned language models are zero-shot learners." *arXiv preprint arXiv:2109.01652*

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In *Advances in Neural Information Processing Systems*. Curran Associates, Inc. [Link]

Wu, Chenfei, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. "Visual chatgpt: Talking, drawing and editing with visual foundation models." *arXiv preprint arXiv:2303.04671*

Wu, Qiucheng, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. 2023. "Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Yang, Cheng-Fu, Wan-Cyuan Fan, Fu-En Yang, and Yu-Chiang Frank Wang. 2021. "Layouttransformer: Scene layout generation with conceptual and spatial diversity." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

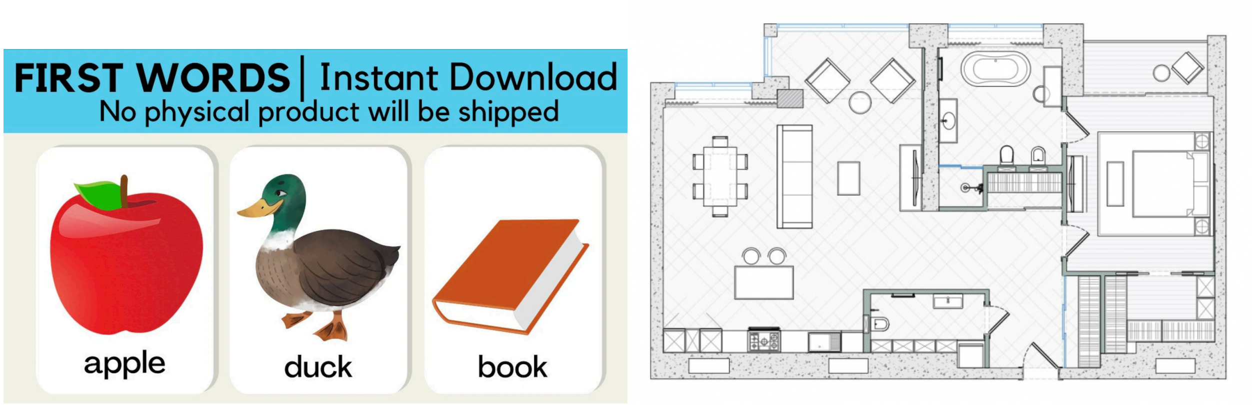
Yang, Zhengyuan, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng et al. 2023. "Reco: Region-controlled text-to-image generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, Yuwei, Zihan Wang, and Jingbo Shang. 2023. "ClusterLLM: Large Language Models as a Guide for Text Clustering." *arXiv preprint arXiv:2305.14871*

Zheng, Xinru, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. 2019. "Content-aware generative modeling of graphic design layouts." *ACM Transactions on Graphics (TOG)* 38 (4): 1–15

Appendices

A.1 Why is Non-Overlapping Important?


FIRST WORDS | Instant Download
No physical product will be shipped
apple
duck
book

(a) Book for Children

(b) 2D Floor Plan Design

Figure A 1: Children's book and 2D floor plan design (Etsy 2024; Archicgi 2019).

Non-overlapping objects are a crucial requirement for many real-world design tasks. For instance, educational books aimed at helping young children learn to read through pictures necessitate clear separation between objects to facilitate easy interpretation and knowledge absorption. Similarly, 2D-floor plan designs demand distinct spaces for each piece of furniture to avoid a cluttered appearance, ensuring that the layout remains accessible and easy to understand for viewers. By enhancing the understanding of prompts that specify non-overlapping objects, we can provide a wealth of inspiration for designers in their practical work.

A.2 Prompts for Layout Generation of LMDpp

We list our LLM prompts in Table A 1 and our in-context examples in Tables A 2 and A 3. In the prompt header, we include instructions on the constraints related to object overlap, spatial relationships, and numeracy. Additionally, we provide explanations on why the given layouts are mathematically sensible in the in-context examples, thereby enabling the LLM to reason and design the most effective layout bounding boxes within its capabilities.

Table A 1: LMDpp prompt header. We use a fixed prompt header for layout generation.

```
You are an intelligent bounding box generator. I will provide you with a description
    of a photo, scene, or painting. Your task is to generate the bounding boxes for
    the objects mentioned in the description and a background prompt describing the
    scene. The images are of size 512x512. The top-left corner has coordinates [0, 0].
    The bottom-right corner has coordinates [512, 512]. The bounding boxes should not
    overlap or go beyond the image boundaries. Each bounding box should be in the
    format of (object name, [top-left x coordinate, top-left y coordinate, box width,
    box height]) and should not include more than one object.
To prevent overlap, for any two boxes (box1 and box2), the condition (x1 + w1 <= x2)
    or (x2 + w2 <= x1) or (y1 + h1 <= y2) or (y2 + h2 <= y1) must be met.
Furthermore, if the description uses spatial keywords (’left’, ’right’, ’above’,
    ’below’), the positioning of the bounding boxes must reflect these relationships
    accurately. This means adjusting the centroids of the boxes (cx1, cy1 for box1;
    cx2, cy2 for box2) so that: cx1 < cx2 if Object A is to the left of B; cx1 > cx2
    if A is to the right of B; cy1 < cy2 if A is above B; and cy1 > cy2 if A is below
    B. Centroids are calculated as cx = x + w//2 and cy = y + h//2.
Objects defined within the bounding boxes should not be repeated in the scene’s
    background description. Exclude any non-relevant or omitted objects from this
    background narrative. Use ”A realistic scene” as the background prompt if no
    background is given in the prompt. If needed, you can make reasonable guesses.
    Please refer to the example below for the desired format.
```

Table A 2: LMDpp in-context example 1. We use fixed in-context examples for layout generation.

```
<Scene One Begins>
Caption: A white background with 3 objects ([’bicycle’, ’boat’, ’laptop’]): the
    bicycle is below the boat; the boat is above the laptop.
Objects: [(’bicycle’, [150, 300, 200, 150]), (’boat’, [150, 150, 200, 100]),
    (’laptop’, [150, 50, 200, 75])]
Background prompt: A white background
Negative prompt:
<Scene One Ends>
<Begin Explanation>
Why does Scene One’s layout have no overlapping and adhere to the caption’s
    description of spatial relationships? Please explain.
For every bounding box (x, y, w, h), max(x + w, y + h) < 512, so that every object is
    within the 512x512 size.
For every two boxes with coordinates (x1, y1, w1, h1) and (x2, y2, w2, h2), the
    condition (x1 + w1 <= x2) or (x2 + w2 <= x1) or (y1 + h1 <= y2) or (y2 + h2 <= y1)
    is met (e.g. for boxes of ’boat’ and ’laptop, x2 + w2 = 150 + 200 = 350 <= 350 =
    x1), so there is no overlapping between each object.
For every two objects A, B with coordinates (xa, ya, wa, ha) and (xb, yb, wb, hb) and
    a spatial relationship Rel(A, B), their spatial relationship is met (e.g. for
    boxes of ’bicycle’ and ’boat’, the centroids for the bicycle is [150+200//2,
    300+150//2] = [250, 375] and the centroid for the boat is [150+200//2, 150+100//2]
    = [250, 200]. Since Rel(bicycle, boat) is “”below and 375 > 200, the two objects
    have the correct spatial relationship as described).
<End Explanation>
```

Table A 3: LMDpp in-context examples continued.

```
<Scene Two Begins>
Caption: A white background with 4 objects ([’elephant’, ’toothbrush’, ’microwave’,
    ’handbag’]): the elephant is below the toothbrush; the toothbrush is to the right
    of the microwave; the microwave is to the right of the handbag.
Objects: [(’elephant’, [150, 300, 200, 150]), (’toothbrush’, [300, 150, 50, 150]),
    (’microwave’, [200, 50, 100, 100]), (’handbag’, [50, 50, 100, 100])]
Background prompt: A white background
Negative prompt:
<Scene Two Ends>
<Begin Explanation>
Why does Scene Two’s layout have no overlapping and adhere to the caption’s
    description of spatial relationships? Please explain.
For every bounding box (x, y, w, h), max(x + w, y + h) < 512, so that every object is
    within the 512x512 size.
For every two boxes with coordinates (x1, y1, w1, h1) and (x2, y2, w2, h2), the
    condition (x1 + w1 <= x2) or (x2 + w2 <= x1) or (y1 + h1 <= y2) or (y2 + h2 <= y1)
    is met (e.g. for boxes of ’elephant’ and ’handbag’, x2 + w2 = 50 + 100 = 150 <=
    150 = x1), so there is no overlapping between each object.
For every two objects A, B with coordinates (xa, ya, wa, ha) and (xb, yb, wb, hb) and
    a spatial relationship Rel(A, B), their spatial relationship is met (e.g. for
    boxes of ’toothbrush’ and ’microwave’, the centroids for the toothbrush are
    [300+150//2, 150+150//2] = [375, 225] and the centroids for the microwave are
    [200+100//2, 50+100//2] = [250, 100]. Since Rel(toothbrush, microwave) is “”right
    and 225 > 100, the two objects have the correct spatial relationship as described).
<End Explanation>
<Scene Three Begins>
Caption: A white background with 5 objects ([’stop sign’, ’sink’, ’clock’, ’tennis
    racket’, ’couch’]): the stop sign is below the sink; the sink is to the left of
    the clock; the clock is below the tennis racket; the tennis racket is to the left
    of the couch.
Objects: [(’stop sign’, [150, 300, 80, 80]), (’sink’, [150, 200, 80, 80]), (’clock’,
    [250, 200, 80, 80]), (’tennis racket’, [250, 100, 80, 80]), (’couch’, [350, 100,
    120, 80])]
Background prompt: A white background
Negative prompt:
<Scene Three Ends>
<Begin Explanation>
Why does Scene Three’s layout have no overlapping and adhere to the caption’s
    description of spatial relationships? Please explain.
For every bounding box (x, y, w, h), max(x + w, y + h) < 512, so that every object is
    within the 512x512 size.
For every two boxes with coordinates (x1, y1, w1, h1) and (x2, y2, w2, h2), the
    condition (x1 + w1 <= x2) or (x2 + w2 <= x1) or (y1 + h1 <= y2) or (y2 + h2 <= y1)
    is met (e.g. for boxes of ’clock’ and ’couch’, y2 + h2 = 100 + 80 = 180 <= 200 =
    y1), so there is no overlapping between each object.
For every two objects A, B with coordinates (xa, ya, wa, ha) and (xb, yb, wb, hb) and
    a spatial relationship Rel(A, B), their spatial relationship is met (e.g. for
    boxes of ’stop sign’ and ’sink’, the centroids for the stop sign is [150+80//2,
    300+80//2] = [190, 340] and the centroid for the sink is [150+80//2, 200+80//2] =
    [190, 240]. Since Rel(stop sign, sink) is “”below and 340 > 240, the two objects
    have the correct spatial relationship as described).
<End Explanation>
```