SLayR: Scene Layout Generation with Rectified Flow

Anonymous authors
Paper under double-blind review

Abstract

We introduce SLayR, **S**cene **Lay**out Generation with **R**ectified flow, a novel transformer-based model for text-to-layout generation, which can integrate into a complete text-to-image pipeline. SLayR addresses a domain in which current text-to-image pipelines struggle: generating scene layouts that are of significant variety and plausibility, when the given prompt is ambiguous and does not provide constraints on the scene. In this setting, SLayR surpasses existing baselines, including LLMs. To accurately evaluate the layout generation, we introduce a new benchmark suite, including numerical metrics and a carefully designed repeatable human-evaluation procedure that assesses the plausibility and variety of images that are generated. We show that our method sets a new state of the art for achieving high plausibility and variety simultaneously, while being at least 3× times smaller in the number of parameters.

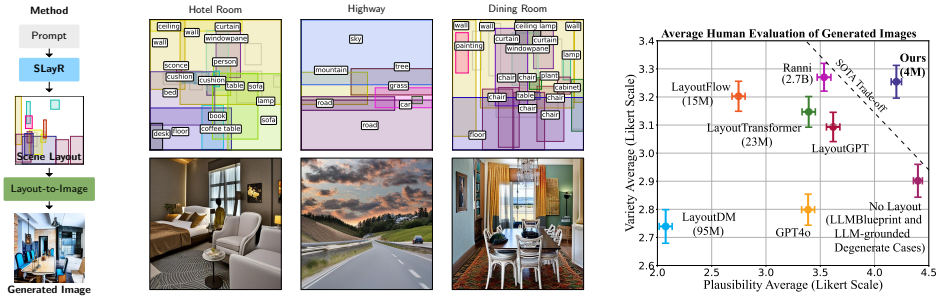

Figure 1: **Left:** We introduce **SLayR**, a method for scene layout generation via rectified flow. **Middle:** SLayR generates scene layouts for unconstrained prompts, which can be rendered using a layout-to-image generator. **Right:** Our method sets a new state of the art in generating more varied and yet plausible scenes than baselines, including LLMs.

1 Introduction

Recent advances in text-to-image modeling have focused on training denoising diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2022) to generate images from a prompt encoding and image noise (Ramesh et al., 2021; Rombach et al., 2022; Saharia et al., 2022; Esser et al., 2024; Zhang et al., 2023a; Sauer et al., 2024), as well as incorporating finer-grained control modalities (Hudson et al., 2023; Kwon et al., 2023; Park et al., 2023; Zhang et al., 2023c; Luo et al., 2024; Shen et al., 2024; Mishra & Subramanyam, 2024; Wu et al., 2023b). Building upon these advancements, prior works have demonstrated the editability and interpretability advantages of a multistage text-to-layout-to-image model, where the user can view and manipulate an intermediate layout consisting of bounding boxes for object-level scene elements (Lian et al., 2024; Feng et al., 2023; Zhou et al., 2024; Gao et al., 2024; Yuan et al., 2024; Öcal et al., 2024; Aguina-Kang et al., 2024). These works use LLMs as text-to-layout generators, and focus on parsing multi-object prompts (e.g. “two dogs next to a cat”). However, a closer inspection reveals that these models do not generate high variety (see Fig. 1, right) or collapse entirely (see Fig. 2), when presented with prompts that have few constraints and leave a high degree of ambiguity. We see this as a critical problem: the models in these cases fail to present knowledge about the structure of scenes as they cannot rely on the

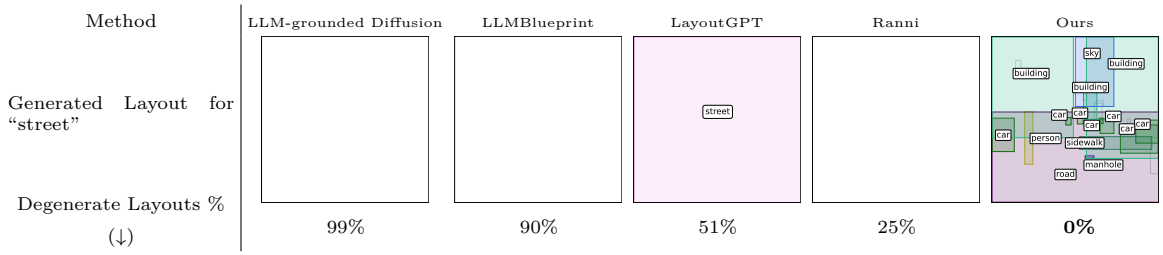

Figure 2: Degenerate layouts (where zero or one trivial bounding box is present) for the prompt "street" from LLM-grounded Diffusion (Lian et al., 2024), LLM Blueprint (Gani et al., 2024), LayoutGPT (Feng et al., 2023), and Ranni (Feng et al., 2024) vs. our layouts. The bottom shows percentages of degenerate layouts from our unconstrained prompt benchmark (See Sec. 4). As visible, LLM approaches for constrained prompts do not generalize to the unconstrained setting.

prompt for specific information.

This motivates us to propose SLayR, a novel lightweight text-to-layout generation model for expanding unconstrained prompts (e.g. "a park", "a beach") into a variety of plausible scene layouts (see Fig. 1, left and middle). Given a CLIP (Radford et al., 2021) embedding of a global scene prompt, we generate the layout using rectified flow (Liu et al., 2022), with a Diffusion Transformer (DiT) (Peebles & Xie, 2023). As unconstrained text-to-layout generation for general images has not been explored before, we assess our layout's plausibility and variety against both LLM-centric baselines and adapted UI/document generation. The experiments show that our method produces a very high variety, while achieving state-of-the-art plausibility in spatial arrangement.

Next, we combine our generated layouts with available layout-to-image generation models (Wang et al., 2024; Li et al., 2023; Xie et al., 2023; Lian et al., 2024) to create a complete text-to-image pipeline. We show that the generated images achieve the highest scores in CMMD (Jayasumana et al., 2024), FID (Heusel et al., 2018), KID (Bińkowski et al., 2021), and HPSv2 (Wu et al., 2023a) compared to the baselines. As true assessment of the image content is only possible by humans, we introduce a comprehensive and repeatable human-evaluation study. The ratings show that our model yields the state-of-the-art trade-off in generating images that are both diverse and plausible. In addition, our pipeline is significantly more lightweight than baselines and can be conditioned on partial layouts and directional constraints, while also providing the ability to edit layouts.

In summary, our contributions are: **1)** we introduce the first model for rectified flow-based text-to-layout generation and show that it produces a large variety of highly plausible layouts for challenging unconstrained prompts, **2)** we establish a well-designed human-evaluation study that can be repeated by others, and **3)** demonstrate that integrating our method into a complete text-to-layout-to-image pipeline yields state-of-the-art in achieving variety and plausibility together. See our supplement to access source code.

2 Related Work

LLMs in Scene Layout Generation. Prior works in 2D layout generation leverage LLMs to parse multi-object prompts into layouts, typically leveraging in-context learning (Lian et al., 2024; Gani et al., 2024; Feng et al., 2023; 2024). Querying these models with unconstrained prompts frequently yields degenerate solutions without meaningful layout information (See Fig. 2). Given that LLM-grounded Diffusion (Lian et al., 2024) and LLM Blueprint (Gani et al., 2024) degenerate in 90% or more cases, we do not evaluate them further. Results on LayoutGPT (Feng et al., 2023) and Ranni (Feng et al., 2024) are provided. To control for the shift to the unconstrained prompt domain, we also adapt the prompt template from (Lian et al., 2024) with in-context examples from our domain, and encouragement of chain-of-thought reasoning (Wei et al., 2023), to meaningfully assess an LLM's capabilities for this task. For the underlying LLM, we use GPT4o (OpenAI et al., 2024)

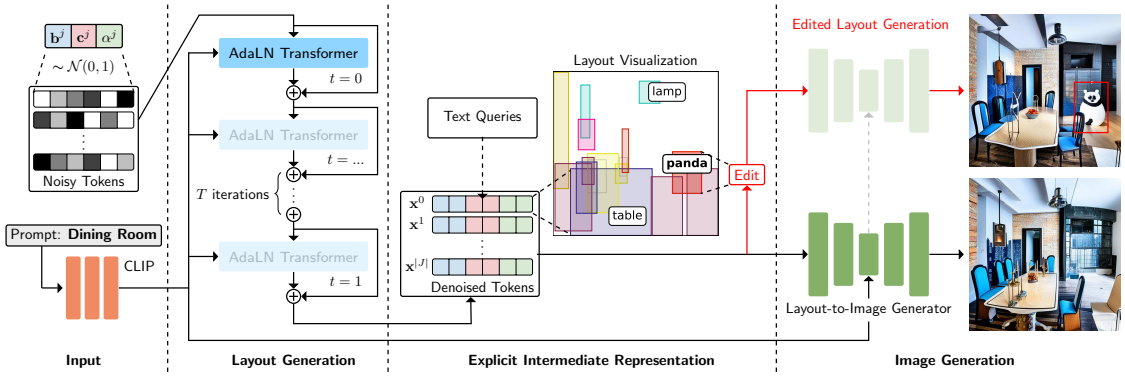

Figure 3: **Method Overview.** Our layout generation model takes a set of noisy tokens and a prompt encoded as a global CLIP embedding as input. The tokens are partitioned into bounding box information $\mathbf{b}^j$, reduced CLIP embeddings $\mathbf{c}^j$, and opacities α^j, with j being the object index. The tokens are then subsequently denoised from $t = 0$ to $t = 1$ using a transformer. For visualization purposes, the user can query the generated layout with labels and edit boxes by adding, moving or removing them. Finally, the generated layout is passed through an off-the-shelf layout-to-image generator.

Adapting UI Generation. Our task of scene layout generation is distinct from User Interface (UI) generation: scene and object captions are from open sets, whereas UI layouts lack global captions and have labels from a small fixed set. Nevertheless, they can serve as interesting baselines, and we adapt several of these models using their conditional generation capabilities. We use LayoutTransformer (Gupta et al., 2021) as a representative for autoregressive transformer approaches, which completes a partial sequence of object bounding boxes to form an image layout. LayoutFormer++ (Jiang et al., 2023) extends LayoutTransformer with added conditioning, but this is not the focus of our assessment of adapted UI generation, and thus it is a redundant baseline. We also adapt LayoutDM (Inoue et al., 2023) and LayoutFlow (Guerreiro et al., 2024) as representative baselines for diffusion-based methods for UI generation (Zhang et al., 2023b; Chai et al., 2023; Levi et al., 2023). For GAN-based approaches (Li et al., 2019), while LayoutGAN++ (Kikuchi et al., 2021) supports inter-bounding-box relationships, the Lagrange multiplier constraint formulation cannot be adapted to support global conditioning. In contrast to our method, UI generation models by design do not extend into the open world scenario.

Rectified Flow. Diffusion modeling has inspired numerous variants and improvements, one of which is rectified flow (Liu et al., 2022). Prior works on the related text-to-image generation task (Liu et al., 2024; Esser et al., 2024). An initial ablation on DDIM (Song et al., 2022), indicates that rectified flow outperforms traditional diffusion approaches (Ho et al., 2020) in this setting. See the supplement for details.

Layout-to-Image Generation. We demonstrate that SLayR integrates well into downstream conditional diffusion models to form a complete text-to-image pipeline, with the added benefits of an interpretable and controllable intermediate layout phase. To control for the effect which the image generator has on the final generated image, we evaluate our layouts across multiple layout-to-image models. Although there are a wide variety of such models, (Chen et al., 2023; Yang et al., 2022; Zhao et al., 2019; Sylvain et al., 2020; Bar-Tal et al., 2023; Xiong et al., 2024) we select four which are publicly available and have been used successfully with LLM-driven layouts (Lian et al., 2024; Feng et al., 2023) or have shown SOTA performance: InstanceDiffusion (Wang et al., 2024), GLIGEN (Li et al., 2023), BoxDiff (Xie et al., 2023), and LMD+ (Lian et al., 2024).

3 Method

The central part of our work is the text-to-layout generation module, which we combine with the existing layout-to-image generators to form a complete text-to-image pipeline. An overview is provided in Fig.3 3, and we explain the details below.

Layout Representation. We start with defining a scene representation as the basis for our generative architecture. A training sample $(\mathbf{x}, P)$ is composed of a global image caption prompt P and a set of J object tokens $\mathbf{x} = \{\mathbf{x}^j \in \mathbb{R}^{d+5}\}_{j \in J}$. The token representation of any single object is composed of

$$\mathbf{x}^j = (\mathbf{b}^j \parallel \mathbf{c}^j \parallel \alpha^j), \tag{1}$$

where $\mathbf{b}^j = (x^j, y^j, w^j, h^j) \in \mathbb{R}^4$ encodes the bounding box coordinates, $\mathbf{c}^j \in \mathbb{R}^d$ is a PCA-reduced CLIP (Radford et al., 2021) embedding, and $\alpha^j \in \mathbb{R}$ is an opacity value that defines the existence of a specific bounding box.

Rectified Flow Preliminaries. We briefly recap the basics of rectified flow introduced in (Liu et al., 2022). Let I be a set of training sample indices and $\{\mathbf{x}_i\}_{i \in I}$ the ground-truth samples whose distribution we would like to learn using our model v. We linearly interpolate between Gaussian noise $\mathbf{x}_i(0)$ and samples $\mathbf{x}_i(1) \equiv \mathbf{x}_i$ across timesteps $t \in [0, 1]$ as follows:

$$\mathbf{x}_i(t) = (1 - t) \cdot \mathbf{x}_i(0) + t \cdot \mathbf{x}_i(1)\,. \tag{2}$$

The model v is trained to take $(\mathbf{x}_i(t), t)$ as input and to predict the derivative of the path between $\mathbf{x}_i(0)$ and $\mathbf{x}_i(1)$, which according to Eq. 2 is $\mathbf{x}_i(1) - \mathbf{x}_i(0)$. The training objective is:

$$\min_v \int_0^1 \mathbb{E}_i[||(\mathbf{x}_i(1) - \mathbf{x}_i(0)) - v(\mathbf{x}_i(t), t)||^2]dt \tag{3}$$

and is optimized with stochastic gradient descent. This optimization is carried out across all available samples of the ground-truth distribution. Following (Liu et al., 2022), noisy values $\mathbf{x}_i(0)$ are resampled at each epoch. The end result is a network v, which is effective at predicting the direction from a noisy sample at an intermediate timestep towards the target distribution. Since a single evaluation may be noisy, the inference is performed by integrating over T timesteps:

$$\mathbf{x}_i(1) = \mathbf{x}_i(0) + \sum_{t=1}^{T} v(\mathbf{x}_i(\frac{t-1}{T}), \frac{t}{T}) \cdot \frac{1}{T}\,. \tag{4}$$

Our Model Architecture. Our rectified flow model is built from multihead AdaLN transformer blocks, which can process tokens $\{\mathbf{x}_i^j\}_{j \in J}$ to iteratively denoise them (Peebles & Xie, 2023). The timestep t, bounding box coordinates $\mathbf{b}_i^j(t)$, and opacity values $\alpha_i^j(t)$ are sinusoidally encoded. The timestep t and a linear projection of the global P_i's CLIP encoding are passed as conditioning of the adaptive layer normalization of the transformer blocks. The tokens represent the objects in the layout and are processed all at once.

Inference begins at $t = 0$ with the set of tokens $\{\mathbf{x}_i^j(t)\}_{j \in J} \equiv \{\mathbf{x}_i^j(0)\}_{j \in J}$ initialized from Gaussian noise. Our model then iteratively processes and updates the tokens from $t = 0$ to $t = 1$ over T iterations using Eq. 4 based on the global prompt conditioning P_i. We project this output back to the dimension of $\mathbf{x}_i^j(t)$ before sinusoidal encoding, in order for the module to serve as the rate of change of $\mathbf{x}_i^j(t)$. A single inference step can be summarized as:

$$\{\mathbf{x}_i^j(t)\}_{j \in J} \leftarrow \{\mathbf{x}_i^j(t - \frac{1}{T})\}_{j \in J} + v(\{\mathbf{x}_i^j(t - \frac{1}{T})\}_{j \in J}, t - \frac{1}{T}, P_i) \cdot \frac{1}{T}\,, \tag{5}$$

Following Eq. 5 until $t = 1$ yields the final layout $\{\mathbf{x}_i^j(1)\}_{j \in J}$ that contains PCA-reduced CLIP embeddings, bounding boxes, and opacities. Tokens with $\alpha_i^j(1) < 0.5$ are considered unused and discarded, please see the supplement for further explanation. For image generation, we unproject each $\mathbf{c}_i^j(1)$ from the PCA space back into the CLIP feature space and pass the embeddings directly into the downstream image generation module.

For visualization of the layouts, we follow the common practice when interpreting visual representations in natural language (Kerr et al., 2023; Qin et al., 2024) and decode CLIP embeddings to text by comparing them to label queries from the user, and selecting the closest query in the embedding space. In the supplement, we

explain the RePaint (Lugmayr et al., 2022; Schröppel et al., 2024) technique for rectified flow to enable *partial layout conditioning*. This enables our model to be guided by partial layouts where only some boxes or labels are given (see Fig. 6). We additionally show how we can impose inter-bounding box positional constraints (i.e., place A to the *left* of B) by adding a directional drift on the bounding boxes during inference. The ability to control our model through these conditions allows it to also work in concert with an LLM to handle complex prompts, where the role of the LLM is to extract the constraints from the prompt, and our method takes care of generating the remaining unspecified scene details.

Training. To construct a training sample from the ground-truth image layout i, we create $\mathbf{c}_i^j$ and $\mathbf{b}_i^j$ for each bounding box j, and initialize α_i^j to 1. To ensure a consistent amount of tokens, we pad the samples by adding tokens with $\alpha_i^j = 0$ and $\mathbf{b}_i^j = 0$, and $\mathbf{c}_i^j$ to the null string embedding. We now treat $\{\mathbf{x}_i^j\}_{j\in J} \equiv \{\mathbf{x}_i^j(1)\}_{j\in J}$, sample $\{\mathbf{x}_i^j(0)\}_{j\in J}$ from Gaussian noise, draw t uniformly from $[0, 1]$, and compute the set of tokens $\{\mathbf{x}_i^j(t)\}_{j\in J}$ by adapting the formula from Eq. 2, which are then passed to the model as input. We refer to the output of the model as $v(\{\mathbf{x}_i^j(t)\}_{j\in J}, t, P_i)$ and compute the training loss derived from Eq. 3:

$$\mathcal{L} = \sum_{i\in I, j\in J} ||\mathbf{x}_i^j(1) - \mathbf{x}_i^j(0) - v(\{\mathbf{x}_i^j(t)\}_{j\in J}, t, P_i)_j||^2 . \tag{6}$$

Human Evaluation. Given the novelty of our problem domain, we argue that human evaluation is most reliable for assessing the plausibility and variety of layouts and therefore introduce a human-evaluation study which can be repeated by others. Assessing human opinions for these criteria directly on layouts is challenging: the evaluators require time to understand the layout diagrams and explain them, and furthermore, assessments are hard to make without actually seeing the image. Following the design principles presented by Otani et al. (2023) in their work on human evaluation of text-to-image generation: 1) *the (evaluation) task should be simple*, and 2) *results should be interpretable*. Therefore, we show participants only images, and omit the underlying image layouts entirely, which may take effort to understand. To make the results interpretable, participants rate these images for their *plausibility* and *variety* on a Likert scale (as recommended in Otani et al. (2023)) from 1 to 5. Image qualities that are assessed in other studies (for example, the overall quality and aesthetic appeal of the image in Liang et al. (2024)) are highly dependent on the conditioned image generator. Therefore, we consider these misleading for our case.

The study is approved by the Ethics Review Board of our institution and complies with local wage regulations. To keep the cost of a survey below 250 USD, we survey 60 participants, who each assess four text-to-layout generation methods at once, each providing ten plausibility questions and ten variety ratings. To increase the stability of the results and test on a larger sample set, each rating is for a collection of three images from the same prompt. The subset of collections, as well as the order they are displayed to the participant, are randomized to control for any potential effects of a fixed ordering. An expanded explanation of our survey design, including the text instructions and screenshots of the survey, can be found in the supplemental material.

4 Experiments

Dataset. We test our method's ability to learn a variety of plausible scene layouts by both training and evaluating on the full ADE20K dataset (Zhou et al., 2018), which contains approximately 27K images and ground-truth layouts for indoor and outdoor scenes, and a rich collection of object arrangements. The sample captions reflect the scene category with no additional constraints (e.g., "beach", "lecture room"). We use the top 30 largest bounding boxes per sample, as this is the default maximum number of bounding boxes supported by InstanceDiffusion (Wang et al., 2024) and we pad samples with fewer bounding boxes. For evaluation, we use the 15 highest represented categories and add in five randomly selected categories to include the dataset's long tail distribution. For each evaluated model, we generate 30 layouts for all 20 selected prompts, and an image conditioned on each layout and corresponding global prompt. The size of this collection of images makes it feasible to assess the results with human evaluation.

Baselines. We compare our method against prior works which are capable of unconstrained layout generation. For LLM-baselines, we evaluate against LayoutGPT (Feng et al., 2023) and Ranni (Feng et al.,

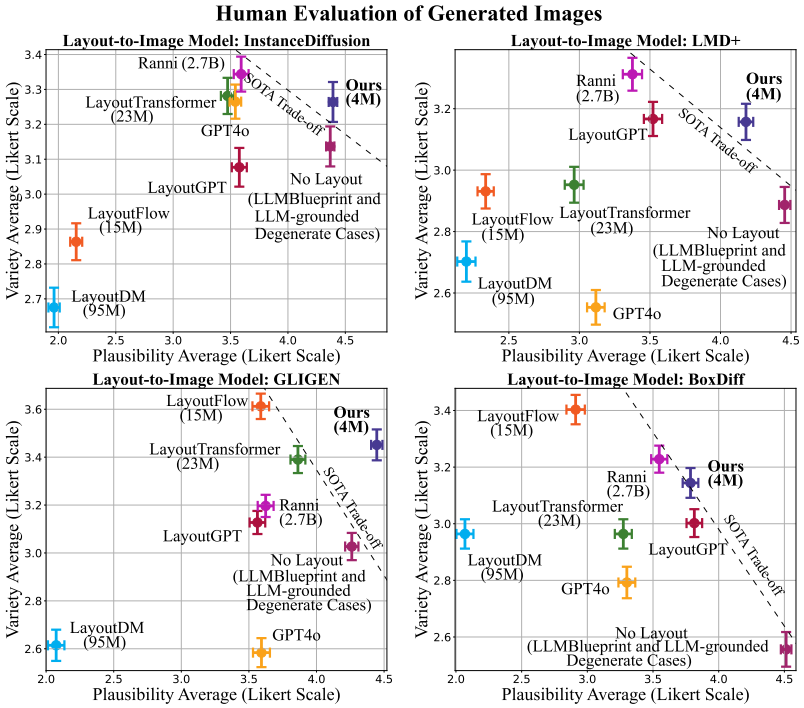

Figure 4: **Human Survey Results.** Our method offers an equal or superior trade-off between plausibility and variety across all measured layout-to-image generators, while being a much smaller model. The error bars indicate standard error.

Model	CMMD (↓)	FID (↓)	KID (10^{-2}) (↓)	HSPv2 (↑)	Image Reward (↑)	VQA (↑)
LayoutFlow	0.25	0.80	0.88	0.23	−1.01	0.80
LayoutDiffusion	0.40	1.08	1.99	0.19	−2.11	0.34
LayoutTransformer	0.06	0.44	0.30	0.23	−1.00	0.75
GPT4o	0.09	0.94	0.45	**0.25**	−0.51	0.88
Ranni	0.07	0.71	0.30	**0.25**	**−0.34**	0.90
LayoutGPT	0.29	2.83	1.76	**0.25**	−0.26	**0.93**
Ours	**0.03**	**0.17**	**0.16**	**0.25**	−0.32	0.88

Table 1: **Image Metrics Comparison.** We evaluate traditional metrics and compare the images generated from layouts of different layout generators. To avoid biases of the image generator, we show the best score among the layout-to-image generators InstanceDiffusion (Wang et al., 2024), GLIGEN (Li et al., 2023), BoxDiff (Xie et al., 2023), and LMD+ (Lian et al., 2024) for each layout generator. Our method achieves strong or state-of-the-art numbers for measured metrics. Although their metrics are strong, Ranni and LayoutGPT are susceptible to degenerate solutions (see Fig. 2)

2024), but discard LLM-grounded Diffusion (Lian et al., 2024) and LLM Blueprint (Gani et al., 2024), as these give degenerate cases in 90%+ of measured cases in our domain (see Fig. 2). To see if LLM performance can be improved with proper in-context examples, we adapt the template from (Lian et al., 2024) with relevant in-context-learning examples from ADE20K. For the underlying LLM, we select the large-scale LLM GPT4o (OpenAI et al., 2024), and refer to this baseline simply as GPT4o. The full template is in the supplement. We test aginst the UI generators LayoutTransformer (Gupta et al., 2021), LayoutDM (Inoue et al., 2023) and LayoutFlow (Guerreiro et al., 2024) by treating the global caption as a scene-wide bounding box and conditioning the model on this bounding box during inference. When training models, we stuck to their respective provided training settings.

Human Evaluation. As shown in Fig. 4, our model achieves a state-of-the-art balance in image plausibility and variety across all measured layout-to-image generators: InstanceDiffusion (Wang et al., 2024),

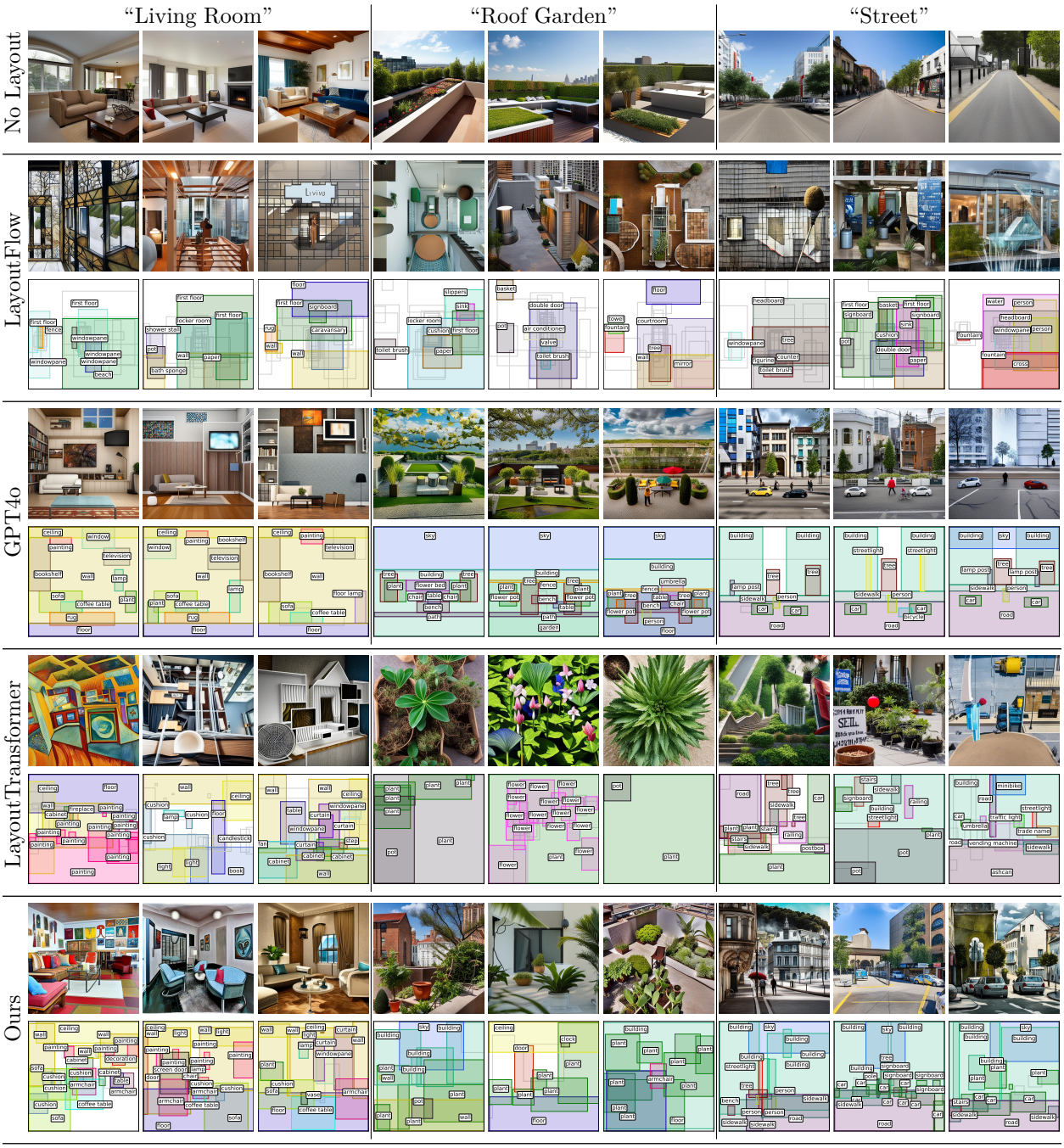

Figure 5: **Qualitative comparison** (Best viewed up close). Layout objects that are depicted in the generated image are highlighted and labeled. From a visual inspection, having no layout produces scenes of little variation in content. LayoutFlow's layouts do not appear to capture scene structure. GPT4o's layouts lack variety. Layout Transformer produces layouts with implausible arrangements of objects, leading to images which do not depict the global prompt accurately. Our method creates plausible and varied layouts, leading to images that are diverse and look realistic. These observations are supported by our human evaluation in Fig. 4. Zoomed-in versions of these layouts for printing are available in the supplemental.

GLIGEN (Li et al., 2023), BoxDiff (Xie et al., 2023), and LMD+ (Lian et al., 2024). The error bars indicate standard error ($s = \frac{\sigma}{\sqrt{n}}$) of the mean human rating, calculated using `numpy` We assume normally distributed

errors. display the approximate number of model parameters added to the full text-to-layout-to-image pipeline by the layout generators that can be locally run. Our model is the smallest by over a factor 3.

Visual Results. We provide a qualitative overview of the generated layouts and the final images in Fig. 5, with InstanceDiffusion (Wang et al., 2024) as the layout-to-image model. We label bounding boxes by querying with all text labels present within ADE20K. From a visual inspection, LayoutTransformer struggles with arranging objects in spatially plausible manner. GPT4o layouts appear somewhat flat, while struggling to make a variety of layouts. Our method appears to produce both plausible and diverse images across a range of global prompts of indoor and outdoor settings.

Generated Image Metrics. We compute established image generation metrics CMMD (Jayasumana et al., 2024), FID (Heusel et al., 2018), KID (Bińkowski et al., 2021), VQA (Lin et al., 2024), HPSv2 (Wu et al., 2023a), and ImageReward (Xu et al., 2023). CMMD, FID and KID compare the distribution of generated images with a ground-truth distribution, while VQA, HSPv2 and ImageReward assess general image quality and alignment with a global caption. Since the conditioned image generator may itself lead to biases in image generation quality, for CMMD, FID, and KID, we establish the ground-truth images by running the layout-to-image generator on the ground-truth layouts. For each layout generator, we display the optimal score over the possible combinations of layout and image generator ((Wang et al., 2024; Li et al., 2023; Xie et al., 2023; Lian et al., 2024)). Images from degenerate layouts from Ranni and LayoutGPT are discarded to more clearly assess the layout's influence. The results are shown in Tab. 1, with state-of-the-art performance in CMMD, FID, KID and HSPv2, and strong results in ImageReward and VQA.

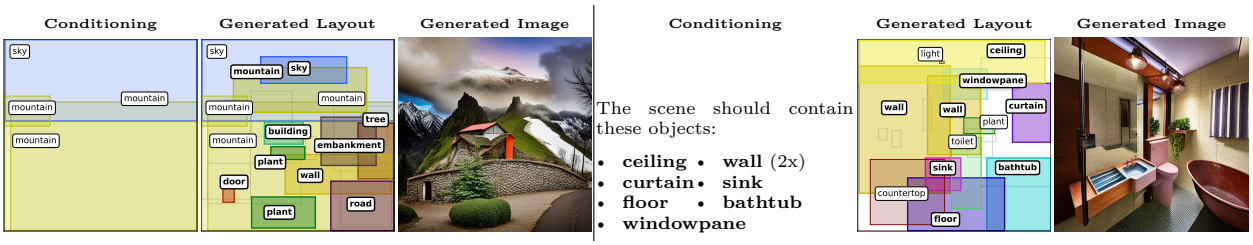

Figure 6: **Disentangled Generation.** Disentangled generation for scenes with the prompt *Snowy Mountain* with a partial layout (**Left**), and *Bathroom* with a *bag of words* (**Right**).

Scene Layout Metrics, and Speed. We consider how to best assess scene layouts for unconstrained prompts. The traditional UI generation metrics of Alignment (Lee et al., 2020) and Overlap (Li et al., 2019) scores are not salient, as real world images often have misaligned or overlapping bounding boxes. Likewise, the layout-FID (Heusel et al., 2018) metric requires a layout-GAN discriminator to compute, which we do not have in this new domain. We compute a standard mIoU (Kikuchi et al., 2021) averaged across sampled scene categories. To provide a more complete evaluation, we introduce metrics aimed to quantify a generated layout's *plausibilty* and *variety* that we describe in full in the supplementary material. We measure the model's generation time on batches of 30 layout samples on an Nvidia A6000 GPU with 32 AMD Ryzen 9 5950X CPUs, 125 GB RAM, except for GPT4o that is accessed through an API. Numerical results are provided in the supplement. Notably, we achieve the highest performance in positional likelihood (how plausibly objects are arranged) and mIoU. Our method ranks second in speed only to LayoutFlow, but we observe no definitive improvement in its layout statistics when the number of inference steps are raised to match our model's speed.

Additional Model Features. We briefly highlight qualities of SLayR which make it appealing to use: In Fig. 6, we show examples of our model's performance in different partial layout generation settings. This feature gives users even more fine-grained control over the image generation process. Additionally, we demonstrate how a text-to-layout-to-image pipeline allows for editing of generated images in Fig. 7. This is accomplished through modifying the intermediate scene layout, and rerunning layout-to-image generator with the original seed and global prompt.

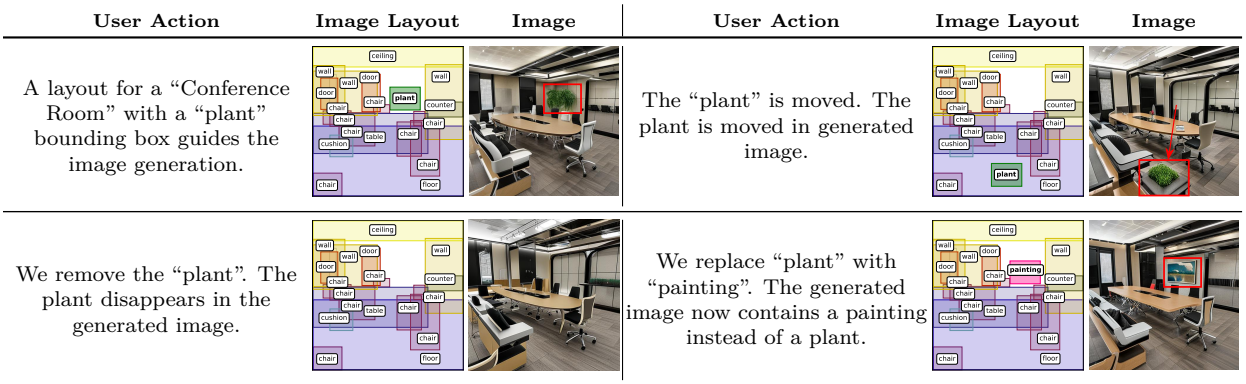

Figure 7: **Editing.** We show how our pipeline enables user editing of images by altering the intermediate scene layout representation. Individual objects can be easily moved, removed, and replaced.

5 Conclusion

We have introduced a text-to-layout model, incorporating it into a text-to-image pipeline with an intermediate and controllable layout representation. With a substantially smaller model, we can generate images with a start-of-the-art balance in plausibility and variety, while achieving high or state-of-the-art performance in generated image quality metrics among competing baselines. In addition, we have introduced a suite of metrics for the new task of scene layout generation, with which we established the foundation to explore image generation pipelines with explicit intermediate layouts in the future.

References

Rio Aguina-Kang, Maxim Gumin, Do Heon Han, Stewart Morris, Seung Jean Yoo, Aditya Ganeshan, R. Kenny Jones, Qiuhong Anna Wei, Kailiang Fu, and Daniel Ritchie. Open-universe indoor scene generation using llm program synthesis and uncurated object databases, 2024. URL https://arxiv.org/abs/2403.09675.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation, 2023. URL https://arxiv.org/abs/2302.08113.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. URL https://arxiv.org/abs/1801.01401.

Shang Chai, Liansheng Zhuang, and Fengying Yan. Layoutdm: Transformer-based diffusion model for layout generation, 2023. URL https://arxiv.org/abs/2305.02567.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance, 2023. URL https://arxiv.org/abs/2304.03373.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Weixi Feng, Wanrong Zhu, Tsu jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models, 2023. URL https://arxiv.org/abs/2305.15393.

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following, 2024. URL https://arxiv.org/abs/2311.17002.

Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts, 2024. URL https://arxiv.org/abs/2310.10640.

Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs, 2024. URL https://arxiv.org/abs/2312.00093.

Julian Jorge Andrade Guerreiro, Naoto Inoue, Kento Masui, Mayu Otani, and Hideki Nakayama. Layoutflow: Flow matching for layout generation, 2024. URL https://arxiv.org/abs/2403.18187.

Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention, 2021. URL https://arxiv.org/abs/2006.14615.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Drew A. Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K. Lampinen, Andrew Jaegle, James L. McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning, 2023. URL https://arxiv.org/abs/2311.17901.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation, 2023. URL https://arxiv.org/abs/2303.08137.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation, 2024. URL https://arxiv.org/abs/2401.09603.

Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. Layoutformer++: Conditional graphic layout generation via constraint serialization and decoding space restriction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18403–18412, 2023.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields, 2023. URL https://arxiv.org/abs/2303.09553.

Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Constrained graphic layout generation via latent optimization. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 88–96. ACM, October 2021. doi: 10.1145/3474085.3475497. URL http://dx.doi.org/10.1145/3474085.3475497.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023. URL https://arxiv.org/abs/2210.10960.

Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. Neural design network: Graphic layout generation with constraints, 2020. URL https://arxiv.org/abs/1912.09421.

Elad Levi, Eli Brosh, Mykola Mykhailych, and Meir Perez. Dlt: Conditioned layout generation with joint discrete-continuous diffusion layout transformer, 2023. URL https://arxiv.org/abs/2303.03755.

Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. Layoutgan: Generating graphic layouts with wireframe discriminators, 2019. URL https://arxiv.org/abs/1901.06767.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation, 2023. URL https://arxiv.org/abs/2301.07093.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models, 2024. URL https://arxiv.org/abs/2305.13655.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation, 2024. URL https://arxiv.org/abs/2312.10240.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation, 2024. URL https://arxiv.org/abs/2404.01291.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation, 2024. URL https://arxiv.org/abs/2309.06380.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022. URL https://arxiv.org/abs/2201.09865.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features, 2024. URL https://arxiv.org/abs/2312.02150.

Rameshwar Mishra and A V Subramanyam. Image synthesis with graph conditioning: Clip-guided diffusion models for scene graphs, 2024. URL https://arxiv.org/abs/2401.14111.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation, 2023. URL https://arxiv.org/abs/2304.01816.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry, 2023. URL https://arxiv.org/abs/2307.12868.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting, 2024. URL https://arxiv.org/abs/2312.16084.

Qualtrics. Qualtrics xm platform, 2024. URL https://www.qualtrics.com. Computer software.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation, 2024. URL https://arxiv.org/abs/2403.12015.

Philipp Schröppel, Christopher Wewer, Jan Eric Lenssen, Eddy Ilg, and Thomas Brox. Neural point cloud diffusion for disentangled 3d shape and appearance generation, 2024. URL https://arxiv.org/abs/2312.14124.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

Guibao Shen, Luozhou Wang, Jiantao Lin, Wenhang Ge, Chaozhe Zhang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Guangyong Chen, Yijun Li, and Ying-Cong Chen. Sg-adapter: Enhancing text-to-image generation with scene graph guidance, 2024. URL https://arxiv.org/abs/2405.15321.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Tristan Sylvain, Pengchuan Zhang, Y. Bengio, R Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts, 03 2020.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation, 2024. URL https://arxiv.org/abs/2402.03290.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023a. URL https://arxiv.org/abs/2306.09341.

Yang Wu, Pengxu Wei, and Liang Lin. Scene graph to image synthesis via knowledge consensus. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):2856–2865, June 2023b. ISSN 2159-5399. doi: 10.1609/aaai.v37i3.25387. URL http://dx.doi.org/10.1609/aaai.v37i3.25387.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion, 2023. URL https://arxiv.org/abs/2307.10816.

Zhexiao Xiong, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, and Nathan Jacobs. Groundingbooth: Grounding text-to-image customization, 2024. URL https://arxiv.org/abs/2409.08520.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2304.05977.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation, 2022. URL https://arxiv.org/abs/2211.15518.

Xuening Yuan, Hongyu Yang, Yueming Zhao, and Di Huang. Dreamscape: 3d scene creation via gaussian splatting joint correlation modeling, 2024. URL https://arxiv.org/abs/2404.09227.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey, 2023a. URL https://arxiv.org/abs/2303.07909.

Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models, 2023b. URL https://arxiv.org/abs/2303.11589.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023c. URL https://arxiv.org/abs/2302.05543.

Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout, 2019. URL https://arxiv.org/abs/1811.11389.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset, 2018. URL https://arxiv.org/abs/1608.05442.

Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting, 2024. URL https://arxiv.org/abs/2402.07207.

Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoglu, and Theo Gevers. Sceneteller: Language-to-3d scene generation, 2024. URL https://arxiv.org/abs/2407.20727.