GLIGEN: Open-Set Grounded Text-to-Image Generation

Yuheng Li[1§], Haotian Liu[1§], Qingyang Wu[2], Fangzhou Mu[1], Jianwei Yang[3], Jianfeng Gao[3], Chunyuan Li[3¶], Yong Jae Lee[1¶]

[1]University of Wisconsin-Madison [2]Columbia University [3]Microsoft

https://gligen.github.io/

Figure 1. GLIGEN enables versatile grounding capabilities for a frozen text-to-image generation model, by feeding different grounding conditions. GLIGEN supports (a) text entity + box, (b) image entity + box, (c) image style and text + box, (d) keypoints, (e) depth map, (f) edge map, (g) normal map, and (h) semantic map.

Abstract

*Large-scale text-to-image diffusion models have made amazing advances. However, the status quo is to use text input alone, which can impede controllability. In this work, we propose GLIGEN, **G**rounded-**L**anguage-to-**I**mage **Gen**eration, a novel approach that builds upon and extends the functionality of existing pre-trained text-to-image diffusion models by enabling them to also be conditioned on grounding inputs. To preserve the vast concept knowledge of the pre-trained model, we freeze all of its weights and inject the grounding information into new trainable layers via a gated mechanism. Our model achieves open-world grounded text2img generation with caption and bounding box condition inputs, and the grounding ability generalizes well to novel spatial configurations and concepts. GLIGEN's zero-shot performance on COCO and LVIS outperforms existing supervised layout-to-image baselines by a large margin.*

§ Part of the work performed at Microsoft; ¶ Co-senior authors

1. Introduction

Image generation research has witnessed huge advances in recent years. Over the past couple of years, GANs [13] were the state-of-the-art, with their latent space and conditional inputs being well-studied for controllable manipulation [42, 54] and generation [25, 27, 41, 75]. Text conditional autoregressive [46, 67] and diffusion [45, 50] models have demonstrated astonishing image quality and concept coverage, due to their more stable learning objectives and large-scale training on web image-text paired data. These models have gained attention even among the general public due to their practical use cases (*e.g*., art design and creation).

Despite exciting progress, existing large-scale text-to-image generation models cannot be conditioned on other input modalities apart from text, and thus lack the ability to precisely localize concepts, use reference images, or other conditional inputs to control the generation process. The current input, *i.e*., natural language alone, restricts the way that information can be expressed. For example, it is difficult to describe the precise location of an object using text, whereas bounding boxes / keypoints can easily achieve this, as shown in Figure 1. While conditional diffusion models [9, 47, 49] and GANs [24, 33, 42, 64] that take in input modalities other than text for inpainting, layout2img generation, *etc*., do exist, they rarely combine those inputs for controllable text2img generation.

Moreover, prior generative models—regardless of the generative model family—are usually independently trained on each task-specific dataset. In contrast, in the recognition field, the long-standing paradigm has been to build recognition models [29, 37, 76] by starting from a foundation model pretrained on large-scale image data [4, 15, 16] or image-text pairs [30, 44, 68]. Since diffusion models have been trained on billions of image-text pairs [47], a natural question is: *Can we build upon existing pretrained diffusion models and endow them with new conditional input modalities?* In this way, analogous to the recognition literature, we may be able to achieve better performance on other generation tasks due to the vast concept knowledge that the pretrained models have, while acquiring more controllability over existing text-to-image generation models.

With the above aims, we propose a method for providing new grounding conditional inputs to pretrained text-to-image diffusion models. As shown in Figure 1, we still retain the text caption as input, but also enable other input modalities such as bounding boxes for grounding concepts, grounding reference images, grounding part keypoints, etc. The key challenge is preserving the original vast concept knowledge in the pretrained model while learning to inject the new grounding information. To prevent knowledge forgetting, we propose to freeze the original model weights and add new trainable gated Transformer layers [61] that take in the new grounding input (*e.g*., bounding box). During training, we gradually fuse the new grounding information into the pretrained model using a gated mechanism [1]. This design enables flexibility in the sampling process during generation for improved quality and controllability; for example, we show that using the full model (all layers) in the first half of the sampling steps and only using the original layers (without the gated Transformer layers) in the latter half can lead to generation results that accurately reflect the grounding conditions while also having high image quality.

In our experiments, we primarily study grounded text2img generation with bounding boxes, inspired by the recent scaling success of learning grounded language-image understanding models with boxes in GLIP [31]. To enable our model to ground open-world vocabulary concepts [29,31,69,72], we use the same pre-trained text encoder (for encoding the caption) to encode each phrase associated with each grounded entity (*i.e*., one phrase per bounding box) and feed the encoded tokens into the newly inserted layers with their encoded location information. Due to the shared text space, we find that our model can generalize to unseen objects even when only trained on the COCO [36] dataset. Its generalization on LVIS [14] outperforms a strong fully-supervised baseline by a large margin. To further improve our model's grounding ability, we unify the object detection and grounding data formats for training, following GLIP [31]. With larger training data, our model's generalization is consistently improved.

Contributions. 1) We propose a new text2img generation method that endows new grounding controllability over existing text2img diffusion models. 2) By preserving the pre-trained weights and learning to gradually integrate the new localization layers, our model achieves open-world grounded text2img generation with bounding box inputs, *i.e*., synthesis of novel localized concepts unobserved in training. 3) Our model's zero-shot performance on layout2img tasks significantly outperforms the prior state-of-the-art, demonstrating the power of building upon large pretrained generative models for downstream tasks.

2. Related Work

Large scale text-to-image generation models. State-of-the-art models in this space are either autoregressive [12, 46, 62, 67] or diffusion [39, 45, 47, 50, 74]. Among autoregressive models, DALL-E [46] is one of the breakthrough works that demonstrates zero-shot abilities, while Parti [67] demonstrates the feasibility of scaling up autoregressive models. Diffusion models have also shown very promising results. DALL-E 2 [45] generates images from the CLIP [44] image space, while Imagen [50] finds the benefit of using pretrained language models. The concurrent Muse [6] demonstrates that masked modeling can achieve SoTA-level generation performance with higher inference speed. However, all of these models usually only take a caption as the input, which can be difficult for conveying other information such as the precise location of an object. Make-A-Scene [12] also incorporates semantic maps into its text-to-image generation, by training an encoder to tokenize semantic masks to condition the generation. However, it can only operate in a closed-set (of 158 categories), whereas our grounded entities can be open-world. A concurrent work eDiff-I [3] shows that by changing the attention map, one can generate objects that roughly follow a semantic map input. However, we believe our interface with boxes is simpler, and more importantly, our method allows other conditioning inputs such as keypoints, which are hard to manipulate through attention.

Image generation from layouts. Given bounding boxes labeled with object categories, the task is to generate a corresponding image [22, 34, 55–57, 65, 71], which is the reverse task of object detection. Layout2Im [71] formulated the problem and combined a VAE object encoder, an LSTM [20] object fuser, and an image decoder to generate the image, using global and object-level adversarial losses [13] to enforce realism and layout correspondence. LostGAN [55, 56] generates a mask representation which is used to normalize features, taking inspiration from StyleGAN [26]. LAMA [34] improves the intermediate mask quality for better image quality. Transformer [60] based methods [22, 65] have also been explored. Critically, existing layout2image methods are closed-set, *i.e*., they can only generate limited localized visual concepts observed in the training set such as the 80 categories in COCO. In contrast, our method represents the first work for *open-set* grounded image generation. A concurrent work ReCo [66] also demonstrates open-set abilities by building upon a pretraned Stable Diffusion model [47]. However, it finetunes the original model weights, which has the potential to lead to knowledge forgetting. Furthermore, it only demonstrates box grounding results whereas we show results on more modalities as shown in the Figure 1.

3. Preliminaries on Latent Diffusion Models

Diffusion-based methods are one of the most effective model families for text2image tasks, among which latent diffusion model (LDM) [47] and its successor Stable Diffusion are the most powerful models publicly available to the research community. To reduce the computational costs of vanilla diffusion model training, LDM proceeds in two stages. The first stage learns a bidirectional mapping network to obtain the latent representation $\boldsymbol{z}$ of the image $\boldsymbol{x}$. The second stage trains a diffusion model on the latent $\boldsymbol{z}$. Since the first stage model produces a fixed bidirectional mapping between $\boldsymbol{x}$ and $\boldsymbol{z}$, from hereon, we focus on the latent generation space of LDM for simplicity.

Training Objective. Starting from noise $\boldsymbol{z}_T$, the model gradually produces less noisy samples $\boldsymbol{z}_{T-1}, \boldsymbol{z}_{T-2}, \cdots, \boldsymbol{z}_0$, conditioned on caption $\boldsymbol{c}$ at every time step t. To learn such

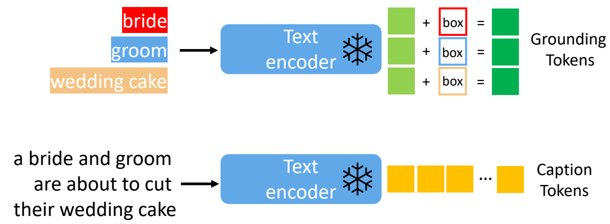

Figure 2. Illustration of grounding token construction process for the bounding box with text case.

a model $f_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$, for each step, the LDM training objective solves the denoising problem on latent representations $\boldsymbol{z}$ of the image $\boldsymbol{x}$:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{LDM}} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\big[\|\boldsymbol{\epsilon} - f_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{c})\|_2^2\big], \quad (1)$$

where t is uniformly sampled from time steps $\{1, \cdots, T\}$, $\boldsymbol{z}_t$ is the step-t noisy variant of input $\boldsymbol{z}$, and $f_{\boldsymbol{\theta}}(*, t, \boldsymbol{c})$ is the $(t, \boldsymbol{c})$-conditioned denoising autoencoder.

Network Architecture. The core of the network architecture is how to encode the conditions, based on which a cleaner version of $\boldsymbol{z}$ is produced. (i) *Denoising Autoencoder*. $f_{\boldsymbol{\theta}}(*, t, \boldsymbol{c})$ is implemented via UNet [48]. It takes in a noisy latent $\boldsymbol{z}$, as well as information from time step t and condition $\boldsymbol{c}$. It consists of a series of ResNet [17] and Transformer [61] blocks. (ii) *Condition Encoding*. In the original LDM, a BERT-like [8] network is trained from scratch to encode each caption into a sequence of text embeddings, $f_{\text{text}}(\boldsymbol{c})$, which is fed into (1) to replace $\boldsymbol{c}$. The caption feature is encoded via a fixed CLIP [44] text encoder in Stable Diffusion. Time t is first mapped to time embedding $\phi(t)$, then injected into the UNet. The caption feature is used in a cross attention layer within each Transformer block. The model learns to predict the noise, following (1).

With large-scale training, the model $f_{\boldsymbol{\theta}}(*, t, \boldsymbol{c})$ is well trained to denoise $\boldsymbol{z}$ based on the caption information only. Though impressive language-to-image generation results have been shown with LDM by pretraining on internet-scale data, it remains challenging to synthesize images where additional grounding input can be instructed, and is thus the focus of our paper.

4. Open-set Grounded Image Generation

4.1. Grounding Instruction Input

For grounded text-to-image generation, there are a variety of ways to ground the generation process via an additional condition. We denote the semantic information of the grounding entity as e, which can be described either through text or an example image; and as $\boldsymbol{l}$ the grounding spatial configuration described with e.g., a bounding box, a set of keypoints, or an edge map, etc. Note that in certain cases, both semantic and spatial information can be represented with $\boldsymbol{l}$ alone (e.g., edge map), in which a single map can represent what objects may be present in the image and where. We define the instruction to a grounded text-to-image model as a composition of the caption and grounded entities:

$$\text{Instruction: } \boldsymbol{y} = (\boldsymbol{c}, \boldsymbol{e}), \quad \text{with} \tag{2}$$

$$\text{Caption: } \boldsymbol{c} = [c_1, \cdots, c_L] \tag{3}$$

$$\text{Grounding: } \boldsymbol{e} = [(e_1, \boldsymbol{l}_1), \cdots, (e_N, \boldsymbol{l}_N)] \tag{4}$$

where L is the caption length, and N is the number of entities to ground. In this work, we primarily study using bounding box as the grounding spatial configuration $\boldsymbol{l}$, because of its large availability and easy annotation for users. For the grounded entity e, we mainly focus on using text as its representation due to simplicity. We process both caption and grounding entities as input tokens to the diffusion model, as described in detail below.

Caption Tokens. The caption $\boldsymbol{c}$ is processed in the same way as in LDM. Specifically, we obtain the caption feature sequence (yellow tokens in Figure 2) using $\boldsymbol{h}^c = [h_1^c, \cdots, h_L^c] = f_{\text{text}}(\boldsymbol{c})$, where h_ℓ^c is the contextualized text feature for the ℓ-th word in the caption.

Grounding Tokens. For each grounded text entity denoted with a bounding box, we represent the location information as $\boldsymbol{l} = [\alpha_{\min}, \beta_{\min}, \alpha_{\max}, \beta_{\max}]$ with its top-left and bottom-right coordinates. For the text entity e, we use the same pre-trained text encoder to obtain its text feature $f_{\text{text}}(e)$ (light green token in Figure 2), and then fuse it with its bounding box information to produce a grounding token (dark green token in Figure 2):

$$h^e = \text{MLP}(f_{\text{text}}(e), \text{Fourier}(\boldsymbol{l})) \tag{5}$$

where Fourier is the Fourier embedding [38], and $\text{MLP}(\cdot, \cdot)$ is a multi-layer perceptron that first concatenates the two inputs across the feature dimension. The grounding token sequence is represented as $\boldsymbol{h}^e = [h_1^e, \cdots, h_N^e]$

From Closed-set to Open-set. Note that existing layout2img works only deal with a closed-set setting (*e.g.*, COCO categories), as they typically learn a vector embedding $\boldsymbol{u}$ per entity, to replace $f_{\text{text}}(e)$ in (5). For a closed-set setting with K concepts, a dictionary of with K embeddings are learned, $\mathbf{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_K]$. While this non-parametric representation works well in the closed-set setting, it has two drawbacks: (1) The conditioning is implemented as a dictionary look-up over $\mathbf{U}$ in the evaluation stage, and thus the model can only ground the observed entities in the generated images, lacking the ability to generalize to ground new entities; (2) No word/phrase is ever utilized in the model condition, and the semantic structure [21] of the underlying language instruction is missing. In contrast, in our open-set design, since the noun entities are processed by the same text encoder that is used to encode the caption, we find that even when the localization information is limited to the concepts in the grounding training datasets, our model can still generalize to other concepts as we will show in our experiments.

Extensions to Other Grounding Conditions. Note that the proposed grounding instruction in Eq (4) is in a general form, though our description thus far has focused on the case of using text as entity e and bounding box as $\boldsymbol{l}$ (the major setting of this paper). To demonstrate the flexibility of the GLIGEN framework, we also study additional representative cases which extend the use scenario of Eq (4).

- *Image Prompt.* While language allows users to describe a rich set of entities in an open-vocabulary manner, sometimes more abstract and fine-grained concepts can be better characterized by example images. To this end, one may describe entity e using an image, instead of language. We use an image encoder to obtain feature $f_{\text{image}}(e)$ which is used in place of $f_{\text{text}}(e)$ in Eq (5) when e is an image.
- *Keypoints.* As a simple parameterization method to specify the spatial configuration of an entity, bounding boxes ease the user-machine interaction interface by providing the height and width of the object layout only. One may consider richer spatial configurations such as keypoints for GLIGEN, by parameterizing $\boldsymbol{l}$ in Eq (4) with a set of keypoint coordinates. Similar to encoding boxes, the Fourier embedding [38] can be applied to each keypoint location $\boldsymbol{l} = [x, y]$.
- *Spatially-aligned conditions.* To enable more fine-grained controlability, spatially-aligned condition maps can be used, such as edge map, depth map, normal map, and semantic map. In these cases, the semantic information e is already contained within each spatial coordinate $\boldsymbol{l}$ of the condition map. A network (*e.g.* conv layers) can be used to encode $\boldsymbol{l}$ into $h \times w$ grounding tokens. We also notice that additionally feeding $\boldsymbol{l}$ into the first conv layer of the UNet can accelerate training. Specifically, the input to the UNet is $\text{CONCAT}(f_l(\boldsymbol{l}), \boldsymbol{z}_t)$ where f_l is a simple downsampling network to reduce $\boldsymbol{l}$ into the same spatial resolution as $\boldsymbol{z}_t$. In this case, the first conv layer of the UNet needs to be trainable.

Figure 1 shows generated examples for these other grounding conditions. Please refer to the supp for more details.

4.2. Continual Learning for Grounded Generation

Our goal is to endow new spatial grounding capabilities to existing large language-to-image generation models. Large diffusion models have been pre-trained on web-scale image-text to gain the required knowledge for synthesizing realistic images based on diverse and complex language instructions. Due to the high pre-training cost and excellent performance, it is important to retain such knowledge in the model weights while expanding the new capability. Hence, we consider to lock the original model weights, and gradually adapt the

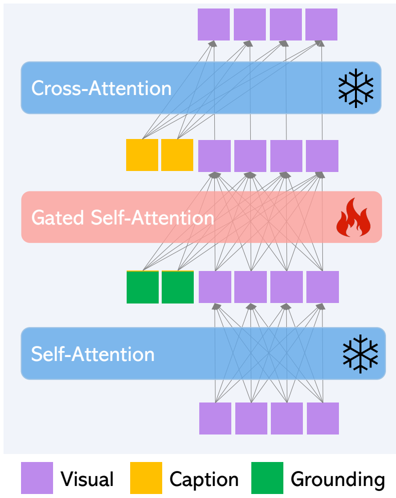

Figure 3. For a pretrained text2img model, the text features are fed into each cross-attention layer. A new gated self-attention layer is inserted to take in the new conditional localization information.

model by tuning new modules.

Gated Self-Attention. We denote $\boldsymbol{v} = [v_1, \cdots, v_M]$ as the visual feature tokens of an image. The original Transformer block of LDM consists of two attention layers: The self-attention over the visual tokens, followed by cross-attention from caption tokens. By considering the residual connection, the two layers can be written:

$$\boldsymbol{v} = \boldsymbol{v} + \text{SelfAttn}(\boldsymbol{v}) \tag{6}$$

$$\boldsymbol{v} = \boldsymbol{v} + \text{CrossAttn}(\boldsymbol{v}, \boldsymbol{h}^c) \tag{7}$$

We freeze these two attention layers and add a new gated self-attention layer to enable the spatial grounding ability; see Figure 3. Specifically, the attention is performed over the concatenation of visual and grounding tokens $[\boldsymbol{v}, \boldsymbol{h}^e]$:

$$\boldsymbol{v} = \boldsymbol{v} + \beta \cdot \tanh(\gamma) \cdot \text{TS}(\text{SelfAttn}([\boldsymbol{v}, \boldsymbol{h}^e])) \tag{8}$$

where $\text{TS}(\cdot)$ is a token selection operation that considers visual tokens only, and γ is a learnable scalar which is initialized as 0. β is set as 1 during the entire training process and is only varied for scheduled sampling during inference (introduced below) for improved quality and controllability. Note that (8) is injected in between (6) and (7). Intuitively, the gated self-attention in (8) allows visual features to leverage bounding box information, and the resulting grounded features are treated as a residual, whose gate is initially set to 0 (due to γ being initialized as 0). This also enables more stable training. Note that a similar idea is used in Flamingo [1]; however, it uses gated cross-attention, which leads to worse performance in our ablation study.

Learning Procedure. We adapt the pre-trained model such that grounding information can be injected while all the original components remain intact. By denoting all the new parameters as $\boldsymbol{\theta}'$, including all gated self-attention layers in (8) and MLP in (2), we use the original denoising objective as in (1) for model continual learning, based on the grounding instruction input $\boldsymbol{y}$:

$$\min_{\boldsymbol{\theta}'} \mathcal{L}_{\text{Grounding}} = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\big[\|\boldsymbol{\epsilon} - f_{\{\boldsymbol{\theta},\boldsymbol{\theta}'\}}(\boldsymbol{z}_t, t, \boldsymbol{y})\|_2^2\big]. \tag{9}$$

Why should the model try to use the new grounding information? Intuitively, predicting the noise that was added to a training image in the reverse diffusion process would be easier if the model could leverage the external knowledge about each object's location. Thus, in this way, the model learns to use the additional localization information while retaining the pre-trained concept knowledge.

Scheduled Sampling in Inference. The standard inference scheme of GLIGEN is to set $\beta = 1$ in (8), and the entire diffusion process is influenced by the grounding tokens. This constant β sampling scheme provides overall good performance in terms of both generation and grounding, but sometimes generates lower quality images compared with the original text2img models (e.g., as Stable Diffusion is finetuned on high aesthetic scored images). To strike a better trade-off between generation and grounding for GLIGEN, we propose a scheduled sampling scheme. As we freeze the original model weights and add new layers to inject new grounding information in training, there is flexibility during inference to schedule the diffusion process to either use both the grounding and language tokens or use only the language tokens of the original model at anytime, by setting different β values in (8). Specifically, we consider a two-stage inference procedure, divided by $\tau \in [0, 1]$. For a diffusion process with T steps, one can set β to 1 at the first $\tau * T$ steps, and set β to 0 for the remaining $(1-\tau) * T$ steps:

$$\beta = \begin{cases} 1, & t \leq \tau * T \quad \text{\# Grounded inference stage} \\ 0, & t > \tau * T \quad \text{\# Standard inference stage} \end{cases} \tag{10}$$

The major benefit of scheduled sampling is improved visual quality as the rough concept location and outline are decided in the early stages, followed by fine-grained details in later stages. It also allows us to extend the model trained in one domain (human keypoint) to other domains (monkey, cartoon characters) as shown in Figure 1.

5. Experiments

We evaluate our model's grounded text2img generation in both the closed-set and open-set settings, and show extensions to other grounding modalities. We conduct our main quantitative experiments by building upon a pretrained LDM on LAION [51], unless stated otherwise.

Model	Generation: FID (↓) Fine-tuned	Generation: FID (↓) Zero-shot	Grounding: YOLO (↑) AP/AP$_{50}$/AP$_{75}$
CogView [10]	-	27.10	-
KNN-Diffusion [2]	-	16.66	-
DALL-E 2 [45]	-	10.39	-
Imagen [50]	-	7.27	-
Re-Imagen [7]	5.25	6.88	
Parti [67]	3.20	7.23	-
LAFITE [75]	8.12	26.94	-
LAFITE2 [73]	4.28	8.42	-
Make-a-Scene [12]	7.55	11.84	-
NÜWA [62]	12.90	-	-
Frido [11]	11.24	-	-
XMC-GAN [70]	9.33	-	-
AttnGAN [63]	35.49	-	-
DF-GAN [59]	21.42	-	-
Obj-GAN [32]	20.75	-	-
LDM [47]	-	12.63	-
LDM*	5.91	11.73	0.6 / 2.0 / 0.3
GLIGEN (COCO2014CD)	5.82	-	21.7 / 39.0 / 21.7
GLIGEN (COCO2014D)	5.61	-	**24.0 / 42.2 / 24.1**
GLIGEN (COCO2014G)	6.38	-	11.2 / 21.2 / 10.7

Table 1. Evaluation of image quality and correspondence to layout on COCO2014 val-set. All numbers are taken from corresponding papers, LDM* is our COCO fine-tuned LDM baseline. Here GLIGEN is built upon LDM.

5.1. Closed-set Grounded Text2Img Generation

We first evaluate the generation quality and grounding accuracy of our model in a closed-set setting. For this, we train and evaluate on the COCO2014 [36] dataset, which is a standard benchmark used in the text2img literature [45, 50, 59, 63, 75], and evaluate how the different types of grounding instructions impact our model's performance.

Grounding instructions. We use the following grounding instructions to train our model: 1) COCO2014D: Detection Data. There are no caption annotations so we use a null caption input [19]. Detection annotations are used as noun-entities. 2) COCO2014CD: Detection + Caption Data. Both caption and detection annotations are used. Note that the noun entities may not always exist in the caption. 3) COCO2014G: Grounding Data. Given the caption annotations, we use GLIP [31], which detects the caption's noun entities in the image, to get pseudo box labels. Please refer to supp for more details about these three types of data.

Baselines. Baseline models are listed in Table 1. Among them, we also finetune an LDM [47] pretrained on LAION 400M [51] on COCO2014 with its caption annotations, which we denote as LDM*. The text2img baselines, as they cannot be conditioned on box inputs, are trained on COCO2014C: Caption Data.

Evaluation metrics. We use the captions and/or box annotations from 30K randomly sampled images to generate 30K images for evaluation. We use *FID* [18] to evaluate image quality. To evaluate grounding accuracy (*i.e.* correspondence between the input bounding box and generated entity), we use the *YOLO score* [35]. Specifically, we use a pretrained YOLO-v4 [5] to detect bounding boxes on the generated images and compare them with the ground truth boxes using average precision (AP). Since prior text2img methods do not support taking box annotations as input, it is not fair to compare with them on this metric. Thus, we only report numbers for the fine-tuned LDM as a reference.

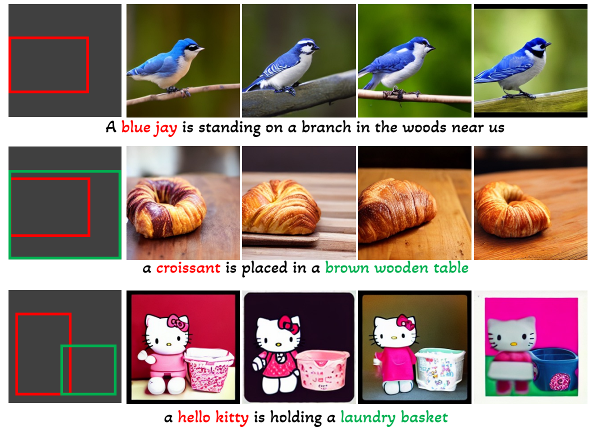

Figure 4. Our model can generalize to open-world concepts even when only trained using localization annotation from COCO.

Results. Table 1 shows the results. First, we see that the image synthesis quality of our approach, as measured by FID, is better than most of the state-of-the-art baselines due to rich visual knowledge learned in the pretraining stage. Next, we find that all three grounding instructions lead to comparable FID to that of the LDM* baseline, which is finetuned on COCO2014 with caption annotations. Our model trained using detection annotation instructions (COCO2014D) has the overall best performance. However, when we evaluate this model on COCO2014CD instructions, we find that it has worse performance (FID: 8.2) – its ability to understand real captions may be limited as it is only trained with the null caption. For the model trained with GLIP grounding instructions (COCO2014G), we actually evaluate it using the COCO2014CD instructions since we need to compute the YOLO score which requires ground-truth detection annotations. Its slightly worse FID may be attributed to its learning from GLIP pseudo-labels. The same reason can explain its low YOLO score (*i.e.*, the model did not see any ground-truth detection annotations during training).

Overall, this experiment shows that: 1) Our model can successfully take in boxes as an additional condition while maintaining image generation quality. 2) All grounding instruction types are useful, which suggests that combining their data together can lead to complementary benefits.

Comparison to Layout2Img generation methods. Thus far, we have seen that our model correctly learns to use the grounding condition. But how accurate is it compared to methods that are specifically designed for layout2img generation? To answer this, we train our model on COCO2017D, which only has detection annotations. We use the 2017 splits (instead of 2014 as before), as it is the standard benchmark

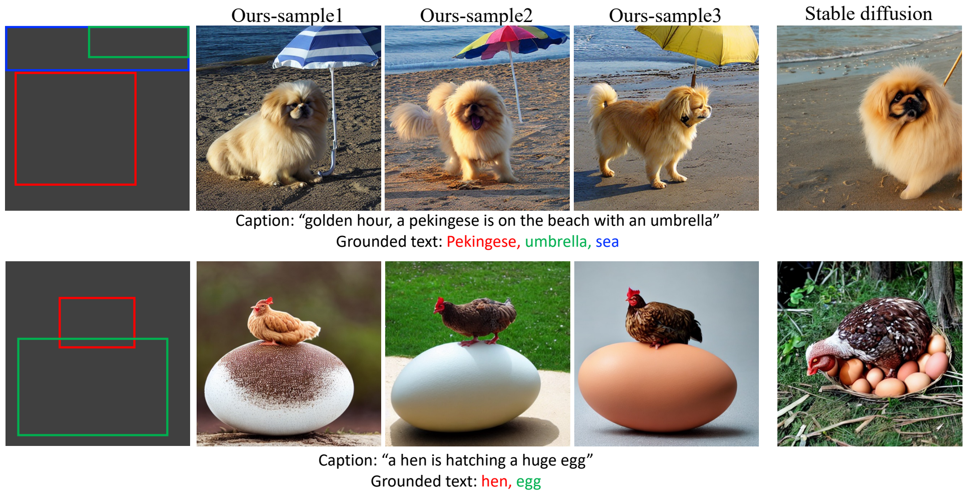

Figure 5. Grounded text2image generation. The baseline lacks grounding ability and can also miss objects *e.g.* "`umbrella`" in a sentence with multiple objects due to CLIP text space, and it also struggles to generate spatially counterfactual concepts.

Model	FID (↓)	YOLO score (AP/AP$_{50}$/AP$_{75}$) (↑)
LostGAN-V2 [56]	42.55	9.1 / 15.3 / 9.8
OCGAN [58]	41.65	-
HCSS [23]	33.68	-
LAMA [35]	31.12	13.40 / 19.70 / 14.90
TwFA [64]	22.15	- / 28.20 / 20.12
GLIGEN-LDM	**21.04**	**22.4 / 36.5 / 24.1**

Table 2. Image quality and correspondence to layout are compared with baselines on COCO2017 val-set.

Model	Training data	AP	AP$_r$	AP$_c$	AP$_f$
LAMA [35]	LVIS	2.0	0.9	1.3	3.2
GLIGEN-LDM	COCO2014CD	6.4	5.8	5.8	7.4
GLIGEN-LDM	COCO2014D	4.4	2.3	3.3	6.5
GLIGEN-LDM	COCO2014G	6.0	4.4	6.1	6.6
GLIGEN-LDM	GoldG,O365	10.6	5.8	9.6	13.8
GLIGEN-LDM	GoldG,O365,SBU,CC3M	11.1	9.0	9.8	13.4
GLIGEN-Stable	GoldG,O365,SBU,CC3M	10.8	8.8	9.9	12.6
Upper-bound	-	25.2	19.0	22.2	31.2

Table 3. GLIP-score on LVIS validation set. Upper-bound is provided by running GLIP on real images scaled to 256×256.

in the layout2img literature. In this experiment, we use the exact same annotation as all layout2img baselines.

Table 2 shows that we achieve the state-of-the-art performance for both image quality and grounding accuracy. We believe the core reason is because previous methods train their model from scratch, whereas we build upon a large-scale pretrained generative model with rich visual semantics. Qualitative comparisons are in the supp.

5.2. Open-set Grounded Text2Img Generation

COCO-training model. We first take GLIGEN trained only with the grounding annotations of COCO (COCO2014CD), and evaluate whether it can generate grounded entities beyond the COCO categories. Figure 4 shows qualitative results, where GLIGEN can ground new concepts such as "`blue jay`", "`croissant`" or ground object attributes such as "`brown wooden table`", beyond the training categories. We hypothesize this is because the gated self-attention of GLIGEN learns to re-position the visual features corresponding to the grounding entities in the caption for the ensuing cross-attention layer, and gains generalization ability due to the shared text spaces in these two layers.

We also quantitatively evaluate our model's zero-shot generation performance on LVIS [14], which contains 1203 long-tail object categories. We use GLIP to predict bounding boxes from the generated images and calculate AP, thus we name it as *GLIP score*. We compare to a state-of-the-art model designed for the layout2img task: LAMA [35]. We train LAMA using the official code on the LVIS training set (in a fully-supervised setting), whereas we directly evaluate our model in a *zero-shot task transfer* manner, by running inference on the LVIS val set without seeing any LVIS labels. Table 3 (first 4 rows) shows the results. Surprisingly, even though our model is only trained on COCO annotations, it outperforms the supervised baseline by a large margin. This is because the baseline, which is trained from scratch, struggles to learn from limited annotations (many of the rare classes in LVIS have fewer than five training samples). In contrast, our model can take advantage of the pretrained model's vast concept knowledge.

Scaling up the training data. We next study our model's open-set capability with much larger training data. Specifically, we follow GLIP [31] and train on Object365 [52] and GoldG [31], which combines two grounding datasets: Flickr [43] and VG [28]. We also use CC3M [53] and SBU [40] with grounding pseudo-labels generated by GLIP.

Table 3 shows the data scaling results. As we scale up the training data, our model's zero-shot performance increases, especially for rare concepts. We also try to finetune the model pretrained on our largest dataset on LVIS and demonstrate its performance in the supp. To demonstrate the generality of our method, we also train our model based on the Stable Diffusion model checkpoint using the largest data. We show some qualitative examples in Figure 5 using this model. Our model gains the grounding ability compared to vanilla Stable Diffusion. We notice that Stable Diffusion model may overlook certain objects ("`umbrella`" in the second example) due to its use of the CLIP text encoder which tends to focus on global scene properties, and may ignore object-level details [3]. It also struggles to generate spatially counterfactual concepts. By explicitly injecting entity information through grounding tokens, our model can improve the grounding ability in two ways: the referred objects are more likely to appear in the generated images, and the objects reside in the specified spatial location.

5.3. Beyond Text Modality Grounding

Image grounded generation. One can also use a reference image to represent a grounded entity as discussed previously. Fig. 1 (b) shows qualitative results, which demonstrate that the visual feature can complement details that are hard to describe by language.

Text and image grounded generation. Besides using either text or image to represent a grounded entity, one can also keep both representations in one model for more creative generation. Fig. 1 (c) shows text grounded generation with style / tone transfer. For the style reference image, we find that grounding it to an image corner or its edge is sufficient. Since the model needs to generate a harmonious style for the entire image, we hypothesize the self-attention layers may broadcast this information to all pixels, thus leading to consistent style for the entire image.

Keypoints grounded generation. We also demonstrate GLIGEN using keypoints for articulate objects control as shown in the Fig. 1 (d). Note that this model is only trained with human keypoint annotations; but it can generalize to other humanoid object due to the scheduled sampling technique we proposed. We also quantitatively study this grounding condition in the supp.

Spatially-aligned condition map grounded generation. Fig. 1 (e-h) demonstrate results for depth map, edge map, normal map, and semantic map grounded generation. These types of conditions allow users to have more fine-grained generation control. See supp for more qualitative results.

5.4. Scheduled Sampling

As stated in Eq. (8) and Eq. (10), we can schedule inference time sampling by setting β to 1 (use extra grounding information) or 0 (reduce to the original pretrained diffusion model). This can make our model exploit different knowledge at different stages.

Fig. 6 qualitatively shows the benefits of our scheduled sampling by setting τ to be 0.2. The images in the same row share the same noise and conditional input. The first row shows that scheduled sampling can be used to improve image quality, as the original Stable Diffusion model is trained with

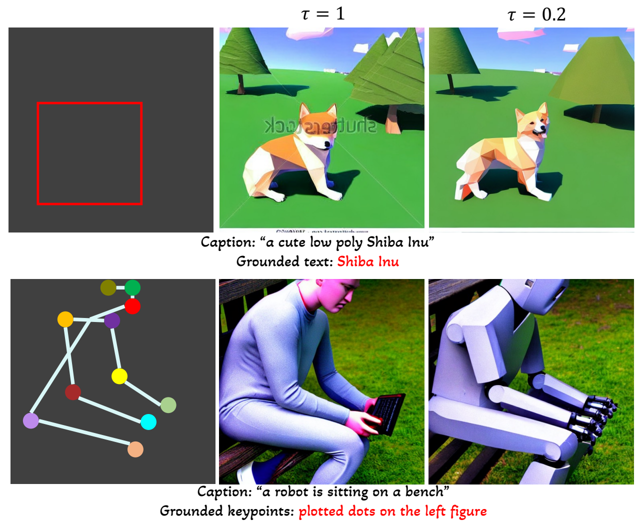

Figure 6. **Scheduled Samping.** It can improve visual or extend a model trained in one domain (e.g., human) to the others.

high quality images. The second row shows a generation example by our model trained with COCO human keypoint annotations. Since this model is purely trained with human keypoints, the final result is biased towards generating a human even if a different object (i.e., robot) is specified in the caption. However, by using scheduled sampling, we can extend this model to generate other objects with a human-like shape.

6. Conclusion

We proposed GLIGEN for expanding pretrained text2img diffusion models with grounding ability, and demonstrated open-world generalization using bounding boxes as the grounding condition. Our method is simple and effective, and can be easily extended to other conditions such as keypoints, reference images, spatially-aligned conditions (e.g., edge map, depth map, etc). One limitation we noticed is that the generated style or aesthetic distribution can shift after adding the new gated self-attention layers (*e.g.*, the model sometimes struggles to generate graphics style images when τ is set to 1), which is probably due to the grounding training data being all natural images. Adding images from more diverse style distributions or further finetuning the model with highly aesthetic images could help alleviate this issue.

Acknowledgement. This work was supported in part by NSF CAREER IIS2150012, NASA 80NSSC21K0295, and Institute of Information & communications Technology Planning & Evaluation(IITP) grants funded by the Korea government(MSIT) (No. 2022- 0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training), and Adobe Data Science Research Award.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022. 2, 5

[2] Oron Ashual, Shelly Sheynin, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 6

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *ArXiv*, abs/2211.01324, 2022. 3, 8

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020. 6

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 6

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 3

[9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021. 2

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 6

[11] Wanshu Fan, Yen-Chun Chen, Dongdong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *ArXiv*, abs/2208.13753, 2022. 6

[12] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *ArXiv*, abs/2203.13131, 2022. 2, 3, 6

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2, 3

[14] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. Lvis: A dataset for large vocabulary instance segmentation. *CVPR*, pages 5351–5359, 2019. 2, 7

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2

[17] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016. 3

[18] 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. In *NIPS*, 2017. 6

[19] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 6

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. 3

[21] Ray S Jackendoff. *Semantic structures*, volume 18. MIT press, 1992. 4

[22] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *ArXiv*, abs/2105.06458, 2021. 3

[23] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *ArXiv*, abs/2105.06458, 2021. 7

[24] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018. 2

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, pages 4396–4405, 2019. 2

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, pages 4396–4405, 2019. 3

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020. 2

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 7

[29] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu,

Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS Track on Datasets and Benchmarks*, 2022. 2

[30] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021. 2

[31] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE, 2022. 2, 6, 7

[32] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019. 6

[33] Yuheng Li, Yijun Li, Jingwan Lu, Eli Shechtman, Yong Jae Lee, and Krishna Kumar Singh. Contrastive learning for diverse disentangled foreground generation. *ArXiv*, abs/2211.02707, 2022. 2

[34] Z. Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. *ICCV*, pages 13799–13808, 2021. 3

[35] Z. Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. *ICCV*, pages 13799–13808, 2021. 6, 7

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 6

[37] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. *CVPR*, 2023. 2

[38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[39] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2

[40] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 7

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *CVPR*, pages 2332–2341, 2019. 2

[42] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *CVPR*, pages 2536–2544, 2016. 2

[43] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015. 7

[44] 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. In *ICML*, 2021. 2, 3

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 2, 6

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 18–24 Jul 2021. 2

[47] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, pages 10674–10685, 2022. 2, 3, 6

[48] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). 3

[49] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2

[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 2, 6

[51] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114, 2021. 5, 6

[52] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. *ICCV*, pages 8429–8438, 2019. 7

[53] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 7

[54] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing.

2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9240–9249, 2020. 2
[55] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. *ICCV*, pages 10530–10539, 2019. 3
[56] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *TPAMI*, 44:5070–5087, 2022. 3, 7
[57] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R. Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. *ArXiv*, abs/2003.07449, 2021. 3
[58] Tristan Sylvain, Pengchuan Zhang, Yoshua Bengio, R. Devon Hjelm, and Shikhar Sharma. Object-centric image generation from layouts. *ArXiv*, abs/2003.07449, 2021. 7
[59] Ming Tao, Hao Tang, Songsong Wu, N. Sebe, Fei Wu, and Xiaoyuan Jing. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *ArXiv*, abs/2008.05865, 2020. 6
[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3
[61] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 2, 3
[62] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pretraining for neural visual world creation. In *European Conference on Computer Vision*, 2022. 2, 6
[63] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018. 6
[64] Zuopeng Yang, Daqing Liu, Chaoyue Wang, J. Yang, and Dacheng Tao. Modeling image composition for complex scene generation. *CVPR*, pages 7754–7763, 2022. 2, 7
[65] Zuopeng Yang, Daqing Liu, Chaoyue Wang, J. Yang, and Dacheng Tao. Modeling image composition for complex scene generation. *CVPR*, pages 7754–7763, 2022. 3
[66] 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. *ArXiv*, abs/2211.15518, 2022. 3
[67] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv*, abs/2206.10789, 2022. 2, 6
[68] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2
[69] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2
[70] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation, 2021. 6
[71] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. *CVPR*, pages 8576–8585, 2019. 3
[72] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2
[73] Yufan Zhou, Chunyuan Li, Changyou Chen, Jianfeng Gao, and Jinhui Xu. Lafite2: Few-shot text-to-image generation. *arXiv preprint arXiv:2210.14124*, 2022. 6
[74] Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, and Jinhui Xu. Shifted diffusion for text-to-image generation. *arXiv preprint arXiv:2211.15388*, 2022. 2
[75] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. *CVPR*, 2022. 2, 6
[76] Xueyan Zou*, Zi-Yi Dou*, Jianwei Yang*, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image and language. *arXiv*, 2022. 2