TOKENCOMPOSE: Text-to-Image Diffusion with Token-level Supervision

Zirui Wang[♣,♠] Zhizhou Sha[♦,♠] Zheng Ding[♠] Yilin Wang[♦,♠] Zhuowen Tu[♠]
[♣]Princeton University [♦]Tsinghua University [♠]University of California, San Diego

https://mlpc-ucsd.github.io/TokenCompose

Figure 1. **Given a user-specified text prompt consisting of object compositions** that are *unlikely* to appear simultaneously in a natural scene, our proposed TOKENCOMPOSE method attains significant performance enhancement over the baseline Latent Diffusion Model (*e.g.,* Stable Diffusion [55]) by being able to generate multiple categories of instances from the prompt more accurately.

Abstract

We present TokenCompose, a Latent Diffusion Model for text-to-image generation that achieves enhanced consistency between user-specified text prompts and model-generated images. Despite its tremendous success, the standard denoising process in the Latent Diffusion Model takes text prompts as conditions only, absent explicit constraint for the consistency between the text prompts and the image contents, leading to unsatisfactory results for composing multiple object categories. Our proposed TokenCompose aims to improve multi-category instance composition by introducing the token-wise consistency terms between the image content and object segmentation maps in the finetuning stage. TokenCompose can be applied directly to the existing training pipeline of text-conditioned diffusion models without extra human labeling information. By finetuning Stable Diffusion with our approach, the model exhibits significant improvements in multi-category instance composition and enhanced photorealism for its generated images. [1]

1. Introduction

Despite the tremendous progress in recent text-to-image diffusion models [7, 17, 21, 52, 55, 57, 58, 72, 75] that have achieved creating images with an increasing level of quality, resolution, photorealism, and diversity, there still exists a major consistency problem between the text prompt and the generated image content. The models loose the composition capability when multiple object categories, especially those not commonly appearing simultaneously in the real world, are included in the text prompt: objects may not appear in the image or their configuration is not pleasantly good looking. Figure 1 shows examples where a state-of-the-art model, Stable Diffusion [55], fails to generate desirable image content from text prompts.

Prototypical generative models of various families [19, 20, 25, 30, 32, 54, 55, 66] have reached maturity. Adding conditional training signals [13–15, 28, 37, 49, 52, 55, 75] to the generative models significantly expands their modeling capability, as well as their scope of application. In the context of Latent Diffusion Models [55], one of the most commonly applied conditions, text (*e.g.,* captions), is injected into layers of the denoising U-Net [56] via cross-attention.

[1]Project done while Zirui Wang, Zhizhou Sha and Yilin Wang interned at UC San Diego. Correspondence to zw1300@cs.princeton.edu

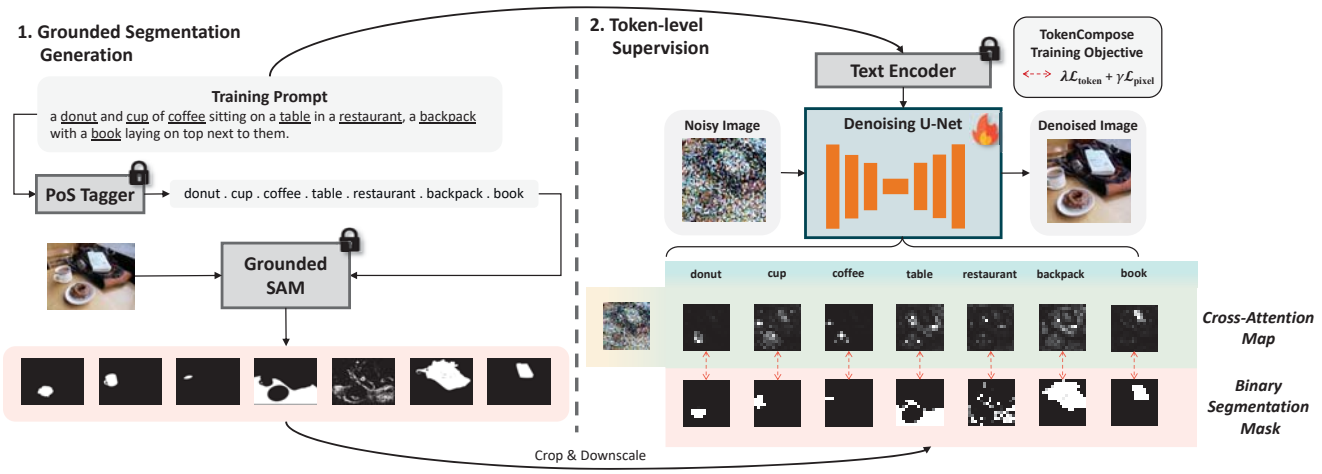

Figure 2. **An overview of the** TOKENCOMPOSE **training pipeline.** Given a training prompt that faithfully describes an image, we adopt a POS tagger [1] and Grounded SAM [31, 42] to extract all binary segmentation maps of the image corresponding to noun tokens from the prompt. Then, we jointly optimize the denoising U-Net of the diffusion model with both its original denoising and our grounding objective.

However, there exists a natural discrepancy between texts (*e.g.,* captions) that are used to train such models and texts (*e.g.,* prompts) that are used for generation. Whereas a caption usually describes a real image faithfully, a prompt can encapsulate image features that do not match the visual scene of any real-world images. Without fine-grained training objectives for its conditioned text, a text-to-image diffusion model often fails to generalize to arbitrary compositions that lie in the prompts [65]. This can be due to the fact that the denoising objective in text-to-image LDM is only optimized to predict the noise given a text prompt, leaving the text condition only as a facilitation in optimizing the denoising function.

By imposing training objectives that operate at the token level in the conditional text, the diffusion model learns what each token from the text means in the context of the image in an *atomistic* manner. Subsequently, it can be better at composing different combinations of words, phrases, *etc*., during inference. However, obtaining the "ground-truth" labels (*e.g.,* segmentation maps) by humans for the corresponding text tokens is label intensive and expensive, especially for the text-image pairs [60] used to train large-scale generative models. Thanks to the recent progress in vision foundation models, Grounding DINO [42] and Segment Anything (SAM) [31], grounding segmentation maps for text tokens can be readily attained automatically.

To this end, we seek to mitigate the composition problem by developing a new algorithm, TOKENCOMPOSE, which leverages models pretrained with image understanding tasks [1, 31, 42] to provide token-level training supervisions to a text-to-image generative model. We show that, by augmenting each noun token from the text prompt of a text-to-image model with segmentation grounding objectives with respect to its respective image during training, the model exhibits significant improvement in object accuracy [18], multi-category instance composition, enhanced photorealism [22] with no additional inference cost for its generated images. Along with our proposed training framework for text-to-image generative models, we also present the MULTIGEN benchmark, which examines the capabilities of a text-to-image generative model to compose multiple categories of instances in a single image.

2. Related Works

Compositional Generation. Efforts that aim to improve compositional image generation for text-conditioned image generative models have been focused on both the training and inference stages. One approach to improving the compositional generation of diffusion models in training is by introducing additional modules, such as a ControlNet, to specify high-level features within the image [75, 76]. However, the modules added to the diffusion models increase the size of the model, leading to additional training and inference costs. Another approach through training is to leverage a reward function to encourage faithful generation of images based on compositional prompts [4, 26, 33, 71]. Albeit their efficacy, reward functions are sparse and do not provide dense supervision signals.

Inference-based methods aim to alter the latent and/or cross-attention maps. Composable Diffusion [41] decomposes a compositional prompt into sub-prompts to generate different latents, and uses a score function to combine the latents together, while Layout Guidance Diffusion [8] uses user-defined tokens and bounding boxes to backpropagate gradients to the latent, and steer the cross-attention maps to focus on specified regions for specified tokens. Other methods apply Gaussian kernels [6] or leverage linguistic features [16, 53] to manipulate the cross-attention map. While these methods do not require further training, they add considerable cost during inference, making it cost $\times 3.37$ times

longer to generate a single image at most, when other generation configurations are kept the same.

Our framework is training-based and does not require additional modules to be incorporated into the image generation pipeline. Furthermore, optimizing attention maps based on segmentation maps provides dense and interpretable supervision. As a training-based method that jointly optimizes token-image correspondence and image generation, the model does not require inference-time manipulations, yet it achieves strong compositionality and competitive image quality for conditional generation.

Benchmarks for Compositional Generation. Several benchmarks have been proposed to evaluate text-conditioned image generative models for compositionality. Most methods evaluate compositionality by binding attributes to and specifying relations between objects. For example, spatial relationships examine whether two different objects appear in the correct spatial layout according to the prompt [2, 11, 18, 26, 58]. Color binding examines whether a text-conditioned image generative model can correctly assign the specified colors to different objects, especially when color assignments are *counterintuitive* [2, 6, 16, 26, 58]. Count binding examines whether a specified instance appears with the right number of counts specified in the prompt [2, 11, 58]. There are additional types of attribute bindings and relation specifications that are evaluated in different benchmarks, such as action and size [2], and shape and texture [26].

However, the majority of such benchmarks confound evaluating capabilities of binding the correct attributes or specifying the correct inter-object relationship with the successful generation of specified objects mentioned in the prompt, making it difficult to evaluate whether improvements are made by stronger attribute assignment & relation specification capabilities or a higher object accuracy [11, 18, 23]. VISOR [18] decouples object accuracy from spatial relationship compositionality by calculating the successful rate of correct spatial relationships conditioned on the successful generation of all specified instances from the prompt. In contrast, almost all other benchmarks do not dissociate these factors and instead evaluate compositionality on a *holistic* basis.

Existing benchmarks on multi-category instance composition focus on successful generation of mostly two categories. On the other hand, leading image generative models have achieved significant improvements in multi-category instance composition [3, 61], which are capable of generating multiple categories of instances with a high success rate. To fill in this research gap in evaluating multi-category instance composition beyond two categories, and to evaluate our training framework in multi-category instance composition, we propose MULTIGEN benchmark, which contains text prompts where each prompt contains objects from an arbitrary combination of multiple categories.

Generative Models for Image Understanding. There has been an upward trend in the use of text-conditioned image generative models for open-vocabulary image understanding tasks, such as classification [12, 34], detection [9, 36], and segmentation [29, 38, 44, 48, 63, 64, 67–70, 77].

An inherent advantage of using generative models for image understanding tasks is being open-vocabulary, as text-to-image diffusion models are trained with open-vocabulary textual prompts. Although results using Stable Diffusion [55] for unsupervised zero-shot image understanding show great potential among methods in the same setting (*i.e.,* zero-shot and open-vocabulary), there still exists a performance gap between these approaches and those that use specialist models for image understanding [5, 10, 31, 42, 74]. This gives us the empirical basis to optimize a generative model (*e.g.,* Stable Diffusion) with knowledge from an understanding model. Furthermore, by optimizing a diffusion model with our approach, we also observe improved performance for downstream segmentation tasks [39, 63], which further reflects successful transfer of this knowledge.

3. Method

3.1. Preliminaries

Diffusion models [25] are widely used in conditional image generative tasks. Given an image $x \in \mathbb{R}^{H\times W\times 3}$, a normally distributed variable ϵ (*e.g.,* noise) is added to the image with a variable extent based on a timestep t. Given a denoising function parameterized by a neural network θ, a noisy image x_t, and a timestep t uniformly sampled from $\{1, \ldots, T\}$, the denoising function learns to predict the noise, ϵ, following the objective (Eq. 1):

$$\mathcal{L}_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \quad (1)$$

To improve the efficiency and control of diffusion models, two changes are performed on the original diffusion recipe – forming the Latent Diffusion Model [55].

First, instead of learning a denoising function in the image space, an image is encoded in a latent state $z_0 = \mathcal{E}(x_0)$ using a variational autoencoder (VAE) [30]. A random noise ϵ is added to the latent z_0, resulting in a noisy latent z_t. The training process involves computing the loss between the predicted noise ϵ_θ and the ground truth noise ϵ to optimize the denoising function.

Second, a conditioning mechanism is added to the denoising function to steer the diffusion process for controllable image generation via cross-attention. In our setting, the condition y is a text prompt that describes the image. To use the text via cross-attention, each token is transformed into an embedding $\tau_\theta(y)$ using a pretrained text encoder [27, 51]. The following shows the denoising objective for an LDM (Eq. 2):

$$\mathcal{L}_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \quad (2)$$

We only optimize the denoising function ϵ_θ, which is parameterized by a U-Net [56] architecture during training. $\mathcal{E}$, $\mathcal{D}$, and τ_θ are kept frozen.

3.2. Token-level Attention Loss

Consider a text prompt that is transformed into text embeddings of length $L_{\tau_\theta(y)}$. As $\mathcal{L}_{LDM}$ only optimizes the function so that it predicts the noise and reconstructs the image latent by removing the noise, the relationship between each token's embedding e_i, $i \in \{1, \dots, L_{\tau_\theta(y)}\}$ and a noisy image latent z_t is not optimized explicitly. This leads to a poor token-level understanding in an LDM, which can be visualized via activations of the multihead cross-attention map (*i.e.,* $\mathcal{A}$) between the token's embedding (*i.e.,* $K \in \mathbb{R}^{H\times L_{\tau_\theta(y)}\times d_k}$), and the noisy image latent (*i.e.,* $Q \in \mathbb{R}^{H\times L_{z_t}\times d_k}$). For each cross-attention layer $m \in M$ with variable latent representation resolutions L_{z_t} in the U-Net, the cross-attention map (*i.e.,* $\mathcal{A} \in \mathbb{R}^{L_{z_t}\times L_{\tau_\theta(y)}}$) is calculated as the following (Eq. 3 and 4):

$$Q^{(h)} = W_Q^{(h)} \cdot \varphi(z_t),\ K^{(h)} = W_K^{(h)} \cdot \tau_\theta(y) \quad (3)$$

$$\mathcal{A} = \frac{1}{H}\sum_h^H \text{softmax}\left(\frac{Q^{(h)}(K^{(h)})^T}{\sqrt{d_k}}\right) \quad (4)$$

where $h \in \{1, \dots, H\}$ represents each head in the multihead cross-attention, φ is a function that flattens a two-dimensional image latent into one dimension, and d_k is the dimension of K.

Empirically, we observe that training a diffusion model with only $\mathcal{L}_{LDM}$ often causes activations of cross attention maps of distinct instance tokens to fail to focus on its corresponding instance appeared in the image during training, which, in turn, results in poor capabilities in composing multiple categories of instances during inference.

To alleviate this issue for better multi-category instance composition, we add a training constraint that supervises activation regions of the cross-attention maps. Specifically, for each text token i that belongs to a noun within the text caption, we acquire the binary segmentation map $\mathcal{M}_i$ from its respective image by leveraging foundation models trained for image understanding [31, 42]. Because cross-attention maps at different layers m of the U-Net have different resolutions, we downscale the resolution of $\mathcal{M}_i$ to match the dimensions of its corresponding $\mathcal{A}_i^{(m)}$ with bilinear interpolation, followed by binarization of all values to form $\mathcal{M}_i^{(m)}$. Different from Layout Diffusion [8], which uses user-defined bounding boxes during inference for gradient-based guidance, we directly apply a loss function $\mathcal{L}_{\text{token}}$ that aggregates activations of cross-attention toward predicted spatial regions $\mathcal{B}_i = \{u \in \mathcal{M}_i \mid u = 1\}$

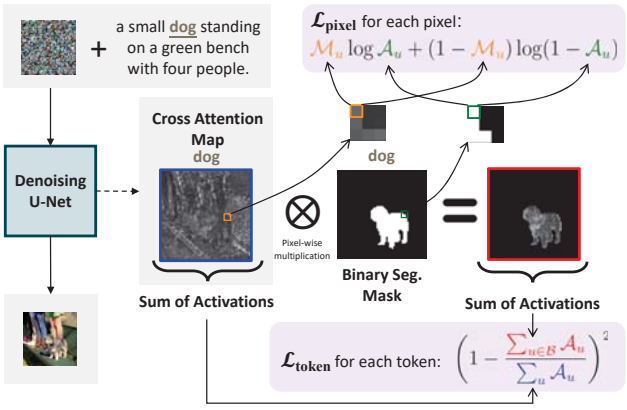

Figure 3. **Illustration of $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$.** We illustrate how $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$ are calculated given a cross-attention map $\mathcal{A}_i$ and a binary segmentation mask $\mathcal{M}_i$. $\mathcal{L}_{\text{token}}$ aggregates attention activations toward non-masked regions of $\mathcal{M}_i$, and this objective is normalized by the total activations of $\mathcal{A}_i$. However, it does *not* constrain where activations should be once inside the non-masked region. $\mathcal{L}_{\text{pixel}}$ gives precise supervision whether a pixel belongs to the segmented region, constraining where activations should be with binary values. However, it is *not* normalized by the total activations of $\mathcal{A}_i$. Combining $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$, we take advantage of the benefit of each objective while minimizing their side effects to a minimum level. We show examples of cross-attention activations from models optimized with $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$, either of them, and neither of them in Figure 4.

jointly with $\mathcal{L}_{LDM}$ during model training. For any layer m, $\mathcal{L}_{\text{token}}$ is defined as follows (Eq. 5):

$$\mathcal{L}_{\text{token}} = \frac{1}{N}\sum_i^N \left(1 - \frac{\sum_{u\in\mathcal{B}_i}^{L_{z_t}} \mathcal{A}_{(i,u)}}{\sum_u^{L_{z_t}} \mathcal{A}_{(i,u)}}\right)^2 \quad (5)$$

where $\mathcal{A}_{(i,u)} \in \mathbb{R}$ represents the scalar attention activation at a spatial location u of $\mathcal{A}_i \in \mathbb{R}^{L_{z_t}}$ for the cross-attention map formed by the latent and the ith token's embedding. Whereas the previous approach [8] calculates the loss on each attention head of a multihead cross-attention module separately, we calculate the loss on the average of cross-attention activations on all heads (see Eq. 4). We find that the latter approach encourages different heads to activate in distinct regions of the cross-attention map, which slightly improves compositional performance and image quality. We add a scaling factor λ to $\mathcal{L}_{\text{token}}$ with respect to $\mathcal{L}_{LDM}$ such that sufficient token-level gradients can be used to optimize token-image consistency while the denoising objective is minimally compromised.

3.3. Pixel-level Attention Loss

Although $\mathcal{L}_{\text{token}}$ substantially aggregates activations of cross-attention maps toward the target regions, a side effect of this aggregation is that the model tends to overly aggregate its activations of the cross-attention map into certain subregions of its target regions. This can be reflected by visually inspecting its cross-attention map during inference

(see Figure 4) and an increased binary cross-entropy loss of cross-attention map activations $\mathcal{A}$ with respect to the target binary segmentation map $\mathcal{M}$. To overcome this problem, we use $\mathcal{L}_{\text{pixel}}$ to counteract. Formally, for a cross-attention map $\mathcal{A}$ in any layer m optimized with $\mathcal{L}_{\text{token}}$, we add the pixel-level cross-entropy objective that is defined as the following (Eq. 6):

$$\mathcal{L}_{\text{pixel}} = -\frac{1}{L_{\tau_\theta(y)} L_{z_t}} \sum_{i}^{L_{\tau_\theta(y)}} \sum_{u}^{L_{z_t}} \big(\mathcal{M}_{(i,u)} \log(\mathcal{A}_{(i,u)}) + \big(1 - \mathcal{M}_{(i,u)}\big) \log\big(1 - \mathcal{A}_{(i,u)}\big)\big) \quad (6)$$

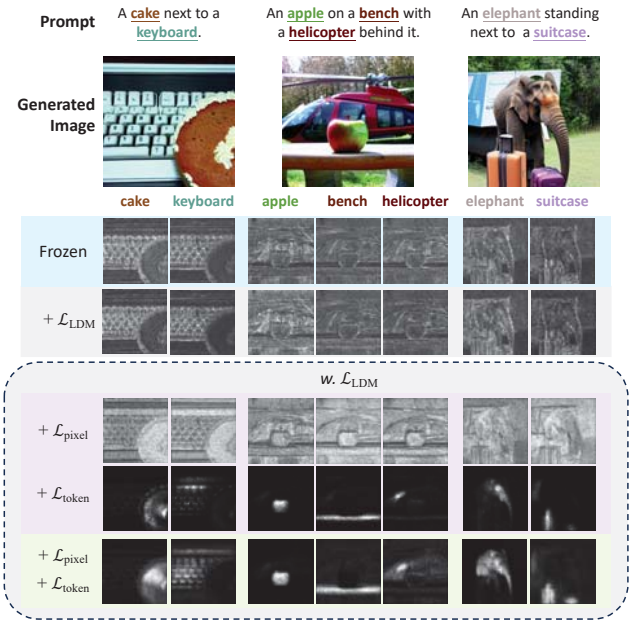

Figure 4. **Impact on cross-attention activations with different objectives.** We firstly demonstrate that finetuning the Stable Diffusion with only $\mathcal{L}_{LDM}$ does not improve grounding capabilities as much. Adding $\mathcal{L}_{\text{pixel}}$ alone causes increased cross-attention activations in general. Adding $\mathcal{L}_{\text{token}}$ plays a vital role in improving token grounding, but leads activations to aggregate in subregions of the targets. By combining $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$, the model shows substantial improvement in grounding text tokens with image features. In this illustration, we apply the null text inversion [47] technique to all models, allowing them to generate the same image for comparable cross-attention maps.

We add a scaling factor γ so that $\mathcal{L}_{\text{pixel}}$ is kept roughly constant while the model is jointly optimized by $\mathcal{L}_{LDM}$ and $\mathcal{L}_{\text{token}}$. We provide a token- and pixel-level optimization illustration in Figure 3 to demonstrate different levels of granularity of $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$. Finally, at any single optimization step, the training objective is as follows (Eq. 7):

$$\mathcal{L}_{\text{TokenCompose}} = \underbrace{\mathcal{L}_{LDM}}_{\text{denoise}} + \sum_{m}^{M} \Bigg(\underbrace{\lambda \mathcal{L}_{\text{token}}^{(m)}}_{\text{token grouding}} + \underbrace{\gamma \mathcal{L}_{\text{pixel}}^{(m)}}_{\text{pixel grounding}} \Bigg) \quad (7)$$

4. Experiment

4.1. Training Details

Dataset. To study the effectiveness of token grounding objectives, we finetune the Stable Diffusion model on a subset of COCO image-caption pairs [39]. Specifically, we first select all unique images from the Visual Spatial Reasoning [40] dataset, as these images have fewer visual-linguistic ambiguities and a greater number of different categories that appear in each image. Then, we use a CLIP model [51] to select the caption with the highest semantic similarity with respect to its corresponding image. Finally, we adopt a pretrained noun parser to parse all nouns from the captions, and leverage Grounded-SAM to generate binary segmentation masks for each noun (or noun phrase) [1, 31, 42]. The final dataset consists of $\sim$ 4526 image-caption pairs and their respective binary segmentation masks. We illustrate a high-level data and training pipeline in Figure 2.

Setup. Our main experiments are performed using Stable Diffusion v1.4 [55], a popular text-to-image diffusion model for high-quality generation. We use both $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$ in addition to its original denoising objective $\mathcal{L}_{LDM}$ with a constant global learning rate of 5e-6 for 24,000 steps (32,000 for Stable Diffusion v2.1) using the AdamW optimizer [43]. We trained the entire U-Net with a batch size of 1 and 4 gradient accumulation steps on a single GPU. We apply center crop to all training images and their respective segmentation maps $\mathcal{M}$. For Stable Diffusion v1.4, the cross-attention layers are located in the U-Net encoder U_{E}, the middle block U_{Mid}, and the decoder U_{D}. We apply $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$ to all cross-attention layers in U_{Mid} and U_{D}.

4.2. Main Results

Baselines. We compare our finetuned Stable Diffusion v1.4 model against several baselines: (1) **Composable Diffusion** [41], which decomposes the prompt into different conditions and uses a score-based mechanism to denoise the image; (2) **Layout Guidance Diffusion** [8], which backpropagates cross-attention map gradients to the noisy latent with user-specified object tokens and bounding boxes for spatially controllable compositional generation; (3) **Structured Diffusion** [16], which automatically parses the prompt into a constituency tree and manipulates cross-attention key and value for compositional generation; and (4) **Attend-and-Excite** [6], which applies Gaussian kernels attention maps from user-specified tokens, and uses smoothed attention maps for compositional generation.

Multi-category Instance Composition. Benchmarks that examine this compositionality focus on the successful generation of multiple categories of instances in the image mentioned in the text condition. We use an existing benchmark, VISOR [18], along with our benchmark, MULTI-GEN, to study the model's capability in multi-category instance composition. The VISOR benchmark obtains all

Method	Multi-category Instance Composition (↑)									Photorealism (↓)		Eff. (↓)
	Object Accuracy	COCO Instances				ADE20K Instances				FID (C)	FID (F)	Latency
		MG2	MG3	MG4	MG5	MG2	MG3	MG4	MG5			
SD [55]	29.86	$90.72_{1.33}$	$50.74_{0.89}$	$11.68_{0.45}$	$0.88_{0.21}$	$89.81_{0.40}$	$53.96_{1.14}$	$16.52_{1.13}$	$1.89_{0.34}$	20.88	71.46	**7.54**$_{0.17}$
Composable [41]	27.83	$63.33_{0.59}$	$21.87_{1.01}$	$3.25_{0.45}$	$0.23_{0.18}$	$69.61_{0.99}$	$29.96_{0.84}$	$6.89_{0.38}$	$0.73_{0.22}$	-	75.57	$13.81_{0.15}$
Layout [8]	43.59	$93.22_{0.69}$	$60.15_{1.58}$	$19.49_{0.88}$	$2.27_{0.44}$	$96.05_{0.34}$	$67.83_{0.90}$	$21.93_{1.34}$	$2.35_{0.41}$	-	74.00	$18.89_{0.20}$
Structured [16]	29.64	$90.40_{1.06}$	$48.64_{1.32}$	$10.71_{0.92}$	$0.68_{0.25}$	$89.25_{0.72}$	$53.05_{1.20}$	$15.76_{0.86}$	$1.74_{0.49}$	21.13	71.68	$7.74_{0.17}$
Attn-Exct [6]	45.13	$93.64_{0.76}$	$65.10_{1.24}$	$28.01_{0.90}$	**6.01**$_{0.61}$	$91.74_{0.49}$	$62.51_{0.94}$	$26.12_{0.78}$	$5.89_{0.40}$	-	71.68	$25.43_{4.89}$
Ours	**52.15**	**98.08**$_{0.40}$	**76.16**$_{1.04}$	**28.81**$_{0.95}$	$3.28_{0.48}$	**97.75**$_{0.34}$	**76.93**$_{1.09}$	**33.92**$_{1.47}$	**6.21**$_{0.62}$	**20.19**	**71.13**	**7.56**$_{0.14}$

Table 1. **Performance of our model in comparison to baselines.** We evaluate the performance based on multi-category instance composition (*i.e.,* Object Accuracy (OA) from VISOR Benchmark [18] and MG2-5 from our MULTIGEN Benchmark), photorealism (*i.e.,* FID [22] from **C**OCO and **F**lickr30K Entities validation splits), and inference efficiency. All comparisons are based on Stable Diffusion 1.4.

unique pairwise combinations of the 80 object categories from COCO [39] and converts each pair (A, B) into a text prompt with an arbitrary spatial relationship (R) following a template "<A> <R> <B>", for example, "a motorcycle to the left of an elephant." With the images generated from such prompts, VISOR uses an open-vocabulary detector [45] to detect the presence and spatial locations of each category in the pair. Object Accuracy (OA) measures the successful rate of generating instances from both categories. VISOR also provides metrics for spatial relationships. However, since our work focuses on multi-category instance composition, we only adopt the OA metric, and report relevant numbers in Table 1.

MULTIGEN uses a similar evaluation strategy compared to VISOR, but is designed to be a more challenging metric for multi-category instance composition. Specifically, given a set of distinct instance categories of size N, we randomly sample 5 categories (*e.g,* A, B, C, D, E), format them into a sentence (*i.e.,* A photo of A, B, C, D, and E.), and use them as the condition for a text-to-image diffusion model to generate the image. Then, we use a strong open-vocabulary detector [46] to detect the presence of these categories in the generated image. We perform the sampling process for 80 categories of COCO instances [39] and 100 categories of ADE20K instances [78, 79] 1,000 times, resulting in 1,000 text prompts as multi-category instance combinations from each dataset. For each generated image, we leverage the detector to detect how many categories of instances appear in the image. We aggregate the overall success rate of generating 2-5 specified categories out of 5 as MG2-5.

Compositional image generation often involves inference variance. To account for this, each prompt in MULTIGEN is used to generate 10 rounds of images, which results in 10 $\times$ 1000 images being generated for each dataset's category combinations. We calculate MG2-5 for each round and report the mean and standard deviation (in $_{\text{subscript}}$) of the MG2-5 success rate out of 10 rounds in Table 1. Based on the evaluation, our model exceeds all baselines in object accuracy and MG2-4. We find that Attend-and-Excite [6] has a considerable success rate in generating all 5 categories, but it falls short of generating 2-4 categories comparably. We conjecture that this is due to the training distribution where captions do not include as many as 5 or more categories of instances, which inevitably leads to diminishing improvements in multi-category instance composition.

Photorealism. We compare the image quality generated from the baselines and our model using the Fréchet Inception Distance (FID) metric [22]. For baselines that do not require additional conditional input other than captions, we calculate the FID score based on 10,000 image-caption pairs sampled from the **C**OCO validation set (**C**). We also report the FID metric based on 1,000 image-caption pairs from the validation set of **F**lickr30K entities (**F**) [50]. As this dataset provides labels and bounding boxes for entities in the captions [73], it can be used by inference-based methods that require such input. We report all applicable scores in Table 1. We also qualitatively evaluate image quality in conjunction with multi-category instance composition across different baselines compared to our model in Figure 5. For fairness, we use the same initial latent in each comparison.

Efficiency. As a training-based method, our model does not require additional inference-time manipulations compared to a standard text-to-image diffusion pipeline. On the other hand, the majority of inference-based methods impose a non-ignorable compute burden for compositional generation, where the slowest baseline, Attend-and-Excite [6], takes more than 3$\times$ of time to generate a single image. We report efficiency results in Table 1 in seconds needed to generate an image on a single NVIDIA RTX 3090 GPU with 50 DDIM steps [62] and classifier-free guidance [24].

4.3. Generalization

We evaluate whether our training method generalizes to different variants of text-to-image models. To this end, we apply $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$ in addition to $\mathcal{L}_{LDM}$ to Stable Diffusion v2.1, and compare the results in multi-category instance composition and photorealism between a frozen baseline and a baseline trained only with $\mathcal{L}_{LDM}$ in Table 3. The results show that these grounding objectives also benefit Stable Diffusion v2.1 remarkably.

4.4. Knowledge Transfer

By learning segmentation maps of each noun token via cross-attention with $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$, the model is expected

	Component	Original	Modified	OA (↑)	COCO Instances (↑) MG3	COCO Instances (↑) MG4	COCO Instances (↑) MG5	ADE20K Instances (↑) MG3	ADE20K Instances (↑) MG4	ADE20K Instances (↑) MG5	FID (C) (↓)
	Ours (Stable Diffusion 1.4)			**52.15**	**76.16**$_{1.04}$	**28.81**$_{0.95}$	**3.28**$_{0.48}$	**76.93**$_{1.09}$	**33.92**$_{1.47}$	**6.21**$_{0.62}$	**20.19**
(i)	finetune	$\mathcal{L}_{LDM} + \lambda\mathcal{L}_{\text{token}} + \gamma\mathcal{L}_{\text{pixel}}$	frozen	29.86	50.74$_{0.89}$	11.68$_{0.45}$	0.88$_{0.21}$	53.96$_{1.14}$	16.52$_{1.13}$	1.89$_{0.34}$	20.88
(ii)	$\lambda\mathcal{L}_{\text{token}} + \gamma\mathcal{L}_{\text{pixel}}$	$\mathcal{L}_{LDM} + \lambda\mathcal{L}_{\text{token}} + \gamma\mathcal{L}_{\text{pixel}}$	$\mathcal{L}_{LDM}$	38.02	63.21$_{1.73}$	19.03$_{1.28}$	1.88$_{0.23}$	62.86$_{1.41}$	22.17$_{1.14}$	3.27$_{0.46}$	23.04
(iii)	$\lambda\mathcal{L}_{\text{token}}$	$\mathcal{L}_{LDM} + \lambda\mathcal{L}_{\text{token}} + \gamma\mathcal{L}_{\text{pixel}}$	$\mathcal{L}_{LDM} + \gamma\mathcal{L}_{\text{pixel}}$	37.46	61.95$_{1.05}$	18.44$_{0.98}$	1.81$_{0.35}$	65.11$_{0.99}$	25.34$_{0.95}$	4.18$_{0.53}$	22.32
(iv)	$\gamma\mathcal{L}_{\text{pixel}}$	$\mathcal{L}_{LDM} + \lambda\mathcal{L}_{\text{token}} + \gamma\mathcal{L}_{\text{pixel}}$	$\mathcal{L}_{LDM} + \lambda\mathcal{L}_{\text{token}}$	49.85	71.61$_{1.06}$	24.94$_{1.24}$	2.83$_{0.62}$	75.37$_{0.90}$	32.91$_{1.53}$	5.58$_{0.46}$	20.60
(v)	layers w. $\lambda\mathcal{L}_{\text{token}} + \gamma\mathcal{L}_{\text{pixel}}$	$U_{\text{Mid}}^{8\times8}, U_{D}^{16\times16}, U_{D}^{32\times32}, U_{D}^{64\times64}$	$U_{E}^{16\times16}, U_{\text{Mid}}^{8\times8}, U_{D}^{16\times16}$	42.92	64.42$_{1.08}$	18.78$_{1.03}$	1.47$_{0.32}$	67.24$_{0.98}$	24.77$_{1.01}$	3.60$_{0.57}$	20.45
			$U_{D}^{16\times16}$	41.66	66.17$_{1.29}$	20.29$_{1.20}$	1.98$_{0.48}$	67.53$_{1.03}$	25.45$_{1.24}$	3.87$_{0.45}$	20.66
			$U_{\text{Mid}}^{8\times8}, U_{D}^{16\times16}$	44.27	65.47$_{1.42}$	19.16$_{1.07}$	1.59$_{0.34}$	69.33$_{1.33}$	26.80$_{1.52}$	4.05$_{0.68}$	20.75
			$U_{\text{Mid}}^{8\times8}, U_{D}^{16\times16}, U_{D}^{32\times32}$	49.89	73.80$_{1.33}$	26.28$_{1.00}$	2.76$_{0.35}$	75.49$_{1.02}$	33.47$_{1.27}$	5.87$_{0.80}$	20.45

Table 2. **Ablation studies**. We show how different objectives and cross-attention layers with $\mathcal{L}_{\text{token}}$ and $\mathcal{L}_{\text{pixel}}$ affect the multi-category instance compositon and photorealism. Qualitative visualizations of the cross-attention maps for (i) - (iv) are in Figure 4.

to gain enhanced abilities to segment instances in the image. To verify that this knowledge is transferred from a segmentation model [31] to a diffusion model, we leverage the COCO-Gen dataset from DAAM [63]. For a fair comparison, we performed null text inversion [47] on all images from this dataset for each model. As the model reconstructs the images, we use DAAM's algorithm and evaluation protocol to calculate the mIoU between the model's cross-attention map and human-annotated segmentation maps. The results are shown in Table 4.

4.5. Downstream Metrics

While TokenCompose is not explicitly optimized for binding attributes such as color, texture, shape to or specifying relations between objects, we show that, by improving the model's capability in multi-category instance composition, we also observe quantitative improvements in downstream compositionality metrics. We evaluate our model using benchmarks proposed by T2I-CompBench [26], which uses various expert models [35, 51, 80] to judge the alignment between compositional prompt and the generated image. We show results of this benchmark in Table 5.

We believe that improvements in these metrics are due to the model having higher chances of composing multiple categories of instances, which serves as a prerequisite for assigning attributes to and relations between objects.

Model	Multi-category Instance Composition (↑) OA	COCO Instances MG3	COCO Instances MG4	COCO Instances MG5	ADE20K Instances MG3	ADE20K Instances MG4	ADE20K Instances MG5	FID (↓) (C)
SD 2.1								
frozen	47.82	70.14	25.57	3.27	75.13	35.07	7.16	19.59
ft. w. $\mathcal{L}_{LDM}$	55.09	76.43	32.07	4.73	77.60	37.09	7.78	20.55
Ours	**60.10**	**80.48**	**36.69**	**5.71**	**79.51**	**39.59**	**8.13**	**19.15**

Table 3. **Model generalization.** We show multi-category instance composition and FID results when we apply our training approach to Stable Diffusion v2.1, which has a different (1) input and cross-attention map resolution, (2) text encoder [27], and (3) training schema (*i.e.,* progressive distillation) [59].

Model	mIoU (↑)
SD 1.4	
frozen	0.5371
ft. w. $\mathcal{L}_{LDM}$	0.5412
Ours	**0.5876**

Table 4. **Enhanced segmentation capabilities.** We examine the grounded segmentaion knowledge transfer from Grounded-SAM [31, 42] to our finetuned diffusion model with DAAM [63].

Model	Attribute Binding (↑) Color	Shape	Texture	Object Relations (↑) Spatial	Non-Spat.	Complex
SD 1.4						
frozen	0.3765	0.3576	0.4642	0.1161	0.3102	0.2795
ft. w. $\mathcal{L}_{LDM}$	0.4647	0.4598	0.5209	0.1326	0.3172	0.2912
Ours	**0.5055**	**0.4852**	**0.5881**	**0.1815**	**0.3173**	**0.2937**

Table 5. **Improvements in downstream metrics.** As successful composition of multiple categories of instances serves as the *foundation* for attribute binding and object relationship specification in compositional generation, our model shows quantitative improvement on relevant metrics, despite it is *not* optimized explicitly for these downstream metrics.

5. Ablations

We ablate (1) incorporating different grounding objectives; and (2) layers applied with grounding objectives in training the denoising U-Net, and evaluate how different design strategies affect multi-category instance composition and photorealism. We show our ablation results in Table 2.

5.1. Grounding Objectives

We compare our model with a model trained only with $\mathcal{L}_{LDM}$ for the same number of optimization steps in (ii). We find that there is a moderate improvement in metrics related to multi-category instance composition comparing to (i), followed by a degeneration in photorealism (*i.e.,* FID). We conjecture that training a model only with $\mathcal{L}_{LDM}$ in the COCO dataset brings an inherent advantage in improving composing multiple categories of instances, as the dataset often contains image-caption pairs where multiple categories of instances appear in the image and are encapsulated by the caption.

We then trained a model with only $\mathcal{L}_{LDM}$ and $\mathcal{L}_{\text{pixel}}$ (iii), and a model trained with only $\mathcal{L}_{LDM}$ and $\mathcal{L}_{\text{token}}$ (iv). We observe that $\mathcal{L}_{\text{pixel}}$ has little effect compared to $\mathcal{L}_{\text{token}}$ when used *alone* with $\mathcal{L}_{LDM}$. $\mathcal{L}_{\text{token}}$ plays a major role in improving multi-category instance composition and photorealism. However, empirically, we find that training the model with

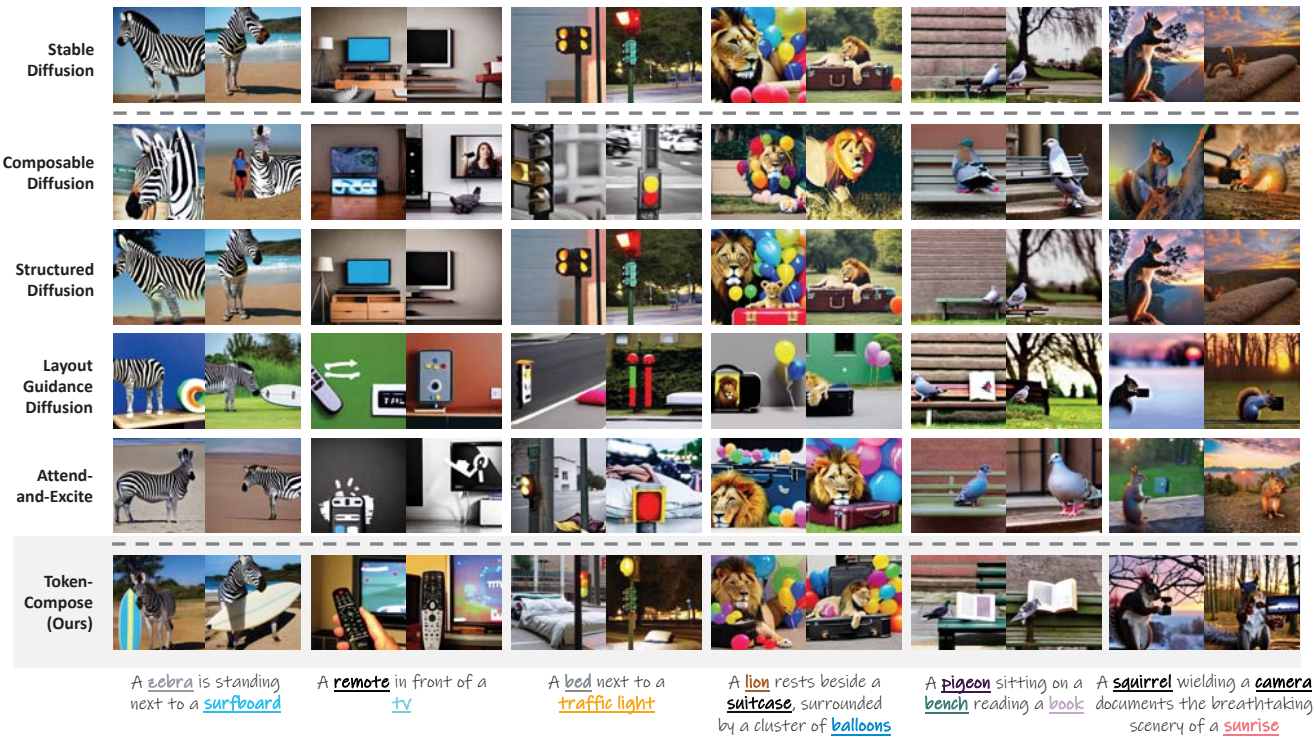

Figure 5. **Qualitative comparison between baselines and our model.** We demonstrate the effectiveness of our training framework in multi-category instance composition compared with a frozen **Stable Diffusion** Model [55], **Composable Diffusion** [41], **Structured Diffusion** [16], **Layout Guidance Diffusion** [8], and **Attend-and-Excite** [6]. The first three columns show composition of two categories that is deemed difficult to be generated from a pretrained Stable Diffusion model (due to rare chances of co-occurrence or significant difference in instance sizes in the real world). The last three columns show the composition of three categories where composing them requires understanding of visual representations of each text token.

only $\mathcal{L}_{LDM}$ and $\mathcal{L}_{\text{token}}$ often leads to attention maps that overly activate in subregions of the instance (see Figure 4). We also observe that different training runs on this combination of losses lead to unstable inference performance in multi-category instance composition.

5.2. Layers with Grounding Objectives

Finally, we experiment with adding $\mathcal{L}_{\text{pixel}}$ and $\mathcal{L}_{\text{token}}$ at different layers of cross-attention of the denoising U-Net (v). We find that adding grounding objectives to the middle block and the decoder of the U-Net improves the overall performance of multi-category instance composition. Removing the constraint from the middle block or adding the constraint to the encoder degrades the performance. Furthermore, for cross-attention layers at the decoder with variable resolutions, we find that the more layers optimized with $\mathcal{L}_{\text{pixel}}$ and $\mathcal{L}_{\text{token}}$, the better the performance in both multi-category instance composition and photorealism.

6. Limitation & Conclusion

Limitation. As one of the pioneering works exploring the potential to improve a text-conditioned generative model with image-token consistency using an understanding model, we only add supervision terms to noun tokens for the text prompt. While we show that this approach improves multi-category instance composition significantly, there are many more elements from the text prompts that one can leverage an understanding model to improve a generative model, such as adjectives, verbs, and/or determiners as fine-grained token-level training objectives.

Conclusion. We explore the possibility of leveraging foundation image understanding models to improve grounding capabilities of a text-conditioned generative model. Our training framework, TOKENCOMPOSE, excels at multi-category instance composition with improved image quality. To facilitate research in this niche, we also propose MULTIGEN, a challenging benchmark that requires a model to generate multiple categories of instances in one image. As a fundamental challenge in compositional generation, we hope our training framework and benchmark can inspire future works that effectively leverage the synergy between understanding and generation to improve either or both directions.

Acknowledgement

This work is supported by NSF Award IIS-2127544. We thank Yifan Xu and Dou Kwark from UC San Diego, Kaiyi Huang from University of Hong Kong, and Adithya Bhaskar and Ofir Press from the Princeton NLP Group for discussions and/or feedback.

References

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019. 2, 5

[2] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023. 3

[3] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions, 2023. 3

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2, 3, 5, 6, 8

[7] Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[8] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 2, 4, 5, 6, 8

[9] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 3

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[11] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023. 3

[12] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. 3

[13] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khác, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 2021. 1

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[15] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 1

[16] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 5, 6, 8

[17] Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, et al. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10135–10145, 2023. 1

[18] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 2, 3, 5, 6

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 1

[20] Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks. *arXiv preprint arXiv:2308.07037*, 2023. 1

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[22] 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. *Advances in neural information processing systems*, 30, 2017. 2, 6

[23] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1552–1565, 2020. 3

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 6

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[26] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 3, 7

[27] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 3, 7

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[29] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*, 2023. 3

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 3

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 3, 4, 5, 7

[32] Kwonjoon Lee, Weijian Xu, Fan Fan, and Zhuowen Tu. Wasserstein introspective neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3702–3711, 2018. 1

[33] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 2

[34] Alexander Cong Li, Mihir Prabhudesai, Shivam Duggal, Ellis Langham Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. 3

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 7

[36] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5567–5577, 2023. 3

[37] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 1

[38] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 3, 5, 6

[40] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 5

[41] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 2, 5, 6, 8

[42] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 4, 5, 7

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[44] Chaofan Ma, Yuhuan Yang, Chen Ju, Fei Zhang, Jinxiang Liu, Yu Wang, Ya Zhang, and Yanfeng Wang. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*, 2023. 3

[45] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 6

[46] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 6

[47] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 5, 7

[48] Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*, 2023. 3

[49] Alexander Quinn 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 *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 1

[50] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Pro-*

ceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015. 6

[51] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 5, 7

[52] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[53] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[54] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1530–1538, Lille, France, 2015. PMLR. 1

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 5, 6, 8

[56] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1, 4

[57] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[58] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3

[59] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 7

[60] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[61] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023. 3

[62] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 6

[63] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, Toronto, Canada, 2023. Association for Computational Linguistics. 3, 7

[64] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469*, 2023. 3

[65] Shengbang Tong, Erik Jones, and Jacob Steinhardt. Mass-producing failures of multimodal systems with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[66] Zhuowen Tu. Learning generative models via discriminative approaches. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1

[67] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 3

[68] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023.

[69] Changming Xiao, Qi Yang, Feng Zhou, and Changshui Zhang. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *arXiv preprint arXiv:2309.04109*, 2023.

[70] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3

[71] 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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[72] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[73] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descrip-

tions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[74] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 3

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2

[76] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023. 2

[77] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3

[78] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 6

[79] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 6

[80] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7571–7580, 2022. 7